ELECTRICAL COMMUNICATION

Technical Journal of the
INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION
and Associate Companies

H. P. WESTMAN, Editor

F. J. MANN, Managing Editor — J. E. SCHLAIKJER, Editorial Assistant

H. T. KOHLHAAS, Consulting Editor

REGIONAL EDITORS

E. G. PORTS, Federal Telephone and Radio Corporation, Newark, New Jersey
B. C. HOLDING, Standard Telephones and Cables, Limited, London, England
P. F. BOURGET, Laboratoire Central de Télécommunications, Paris, France
H. B. WOOD, Standard Telephones and Cables Pty. Limited, Sydney, Australia

EDITORIAL BOARD

H. Busignies, H. H. Buttner, G. Deakin, E. M. Deloraine, W. T. Gibson, Sir Frank Gill, W. Hatton, E. Labin, E. S. McLarn, A. W. Montgomery, Haraden Pratt, G. Rabuteau, F. X. Rettenmeyer, T. R. Scott, C. E. Strong, A. E. Thompson, E. N. Wendell, W. K. Weston

Published Quarterly by the

INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

67 BROAD STREET, NEW YORK 4, N.Y., U.S.A.

Sosthenes Behn, Chairman and President — Charles D. Hilles, Jr., Vice President and Secretary

Subscription, $2.00 per year; single copies, 50 cents

Electrical Communication is indexed in Industrial Arts Index

Copyrighted 1948 by International Telephone and Telegraph Corporation

Volume 25 — June, 1948 — Number 2

CONTENTS

PAGE

COPENHAGEN BROADCASTING HOUSE 107
By Frederik Heegaard and Eigil Cohrt

ZURICH AUTOMATIC TANDEM TOLL EXCHANGE 113
By G. Klingelfuss

EMERGENCY OPERATION OF SOJ-12 CARRIER TELEPHONE SYSTEMS 121
By J. M. Clara and E. Serra Calabuig

NICKEL-IRON ALLOY DUST CORES 126
By S. E. Buckley

VERY-HIGH-FREQUENCY SINGLE-CHANNEL RECEIVER 132
By W. C. Lane and T. C. Clark

HIGH GAIN WITH DISCONE ANTENNAS 139
By A. G. Kandoian, W. Sichak, and R. A. Felsenheld

MODERN DEMONSTRATION OF MACDONALD'S EQUIVALENCE THEOREM 148
By A. G. Clavier

$TM_{0,1}$ MODE IN CIRCULAR WAVE GUIDES WITH TWO COAXIAL DIELECTRICS 152
By Sidney Frankel

HIGH-RATIO MULTIVIBRATOR FREQUENCY DIVIDER 160
By M. Silver and A. Shadowitz

THERMO-ELECTRIC AND CONDUCTIVE PROPERTIES OF BLUE TITANIUM DIOXIDE .. 163
By H. K. Henisch

RELATIONSHIP BETWEEN RATE OF TRANSMISSION OF INFORMATION, FREQUENCY BANDWIDTH, AND SIGNAL-TO-NOISE RATIO 178
By C. W. Earp

NEW DEVELOPMENTS IN MARINE RADIO DIRECTION FINDERS 196
By H. Busignies

H. T. KOHLHAAS, EDITOR, RETIRES 204

CONTRIBUTORS TO THIS ISSUE 205

The Great Concert Studio of Copenhagen Broadcasting House. The stage will accommodate an orchestra and choir of 100 persons each.

Copenhagen Broadcasting House

By FREDERIK HEEGAARD

Post and Telegraph Administration, Copenhagen, Denmark,

and EIGIL COHRT

Standard Electric A/S, Copenhagen, Denmark

COPENHAGEN BROADCASTING HOUSE was built and put in service during the war. In planning the buildings, careful attention was given to obtaining adequate sound insulation among the studios. This was obtained by building each studio as a separate structure with its own foundation. In addition, Broadcasting House is near the center of town and noises from the streets surrounding it must not penetrate to the studios. These street noises are reduced by screening the studios with the office and administration buildings, the concert hall, and by other structures. This arrangement is evident from Fig. 1.

1. Studios

The studio block is divided through the middle by a wide hall with bays, arranged as a lobby for the artists. From the lobby, a studio is entered by passing through at least two sound-proof doors connected by an acoustically heavily damped corridor, which acts as a sound-trap. Associated with each studio is a control room intended for supervision and control of broadcasts and rehearsals in the studio; access to this room is also obtained from the sound-trap. Large triple glass windows provide visibility between the studios and their control rooms without sound leakage. In several cases, a second studio control room is common to two studios, thereby making it possible to operate them as one. Studios 2 and 3 are an example of this. Table I gives data on studios 2 to 8.

In addition, there are a concert hall and three small lecture studios. The former, which is shown in the frontispiece, will accommodate 100 musicians, a choir of 100, and about 1200 auditors. Its volume is 12,000 cubic meters (423,700 cubic feet). The auditors may enter directly from the street. The lecture studios are in the administration block with direct access for the lecturers from the main hall. To obtain the necessary sound insulation, these three studios are provided with floating floors, walls, and ceilings. Two echo rooms of different characteristics are in the studio block.

The technical department is located centrally above the studio hall; it consists of a large amplifier room with the amplifiers mounted in two rows of steel cabinets, a room containing the switching system for connections to the studios, three main control rooms, four sound-recording rooms, and a couple of offices for the operating management. This central position of the technical installations is convenient in view of the fairly complicated cabling between the studios, studio control, main control, amplifier, recording, and relay rooms.

2. Program Circuits

All program circuits have been laid as single-pair paper-insulated cables with a special screen insulated from the lead cover. To decrease cross talk, the pairs have been twisted with two

TABLE I

STUDIO CHARACTERISTICS

Studio	Service	Volume		Reverberation Time in Seconds at 800 Cycles
		Cubic Meters	Cubic Feet	
2	Orchestra, Medium Size	2480	87,669	1.1
3	Orchestra, Small	1450	51,200	Variable
4	Soloist, Chamber Music	475	16,782	0.7
5	Discussions, Small Cabaret with Audience	730	25,776	0.8
6	Soloist, Plays	414	14,618	0.5
7	Plays	1050	47,076	0.5
8	Speech in Plays	47	1,660	0.5

The Great Concert Studio of Copenhagen Broadcasting House. The stage will accommodate an orchestra and choir of 100 persons each.

different pitches, one for cables carrying high signal levels and the other for low-level cables. A view of the amplifier room is shown in Fig. 2.

The amplifier equipment may be divided into the following main groups: (A) program sources, (B) switching and signaling systems, (C) program channels, (D) output patching panel, (E) control apparatus, (F) sound recording system, and (G) talk-back equipment.

A program source may consist of a studio or a combination of studios with their microphones and amplifier system, a telephone line transmitting an outside program, or a playback apparatus for recorded programs. In each studio, a number of microphones may be used to fit the needs of the particular type of program. Each microphone is connected to an *A* amplifier, all of which are located in the amplifier room. The *A* amplifiers are one-stage push-pull units, three of which are mounted on a single panel and share a common power supply as may be seen in Fig. 3. The signal from the *A* amplifier then passes through a bridged-T mixing pad mounted on a desk, shown in Fig. 4, in the studio control room. Next in the circuit is a *B* amplifier. This is a three-stage resistance-coupled amplifier with negative feedback, containing its own power supply. Then follows the main regulating fader, also mounted on the studio control desk, which is connected through

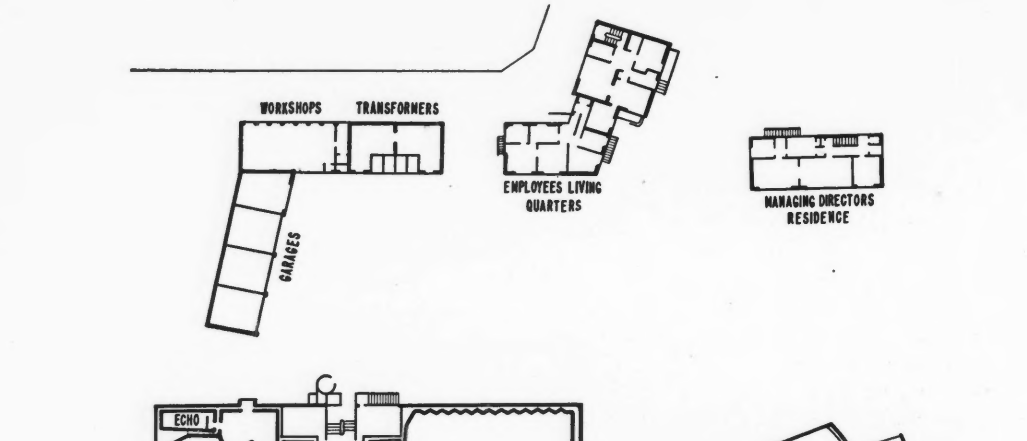


Fig. 1—Plan view of Copenhagen Broadcasting House. The central studio building is screened extensively from city noises by the various administrative and other buildings surrounding it.

a transformer to an appropriate switching system.

The lines carrying outside programs to Broadcasting House arrive in two cables, which also contain the outgoing lines to the transmitters and to the Copenhagen repeater station. The outside-broadcast lines are connected through line transformers, variable equalizers, attenuators, and, if required, *A* amplifiers, before going to the switching points.

Fig. 2—Part of the amplifier room.

In principle, the play-back apparatus in the recording rooms and elsewhere are connected to the switching system in the same way as described for the studios. In this instance, *A* amplifiers have been omitted, the level being higher. They are provided with the necessary equalizing networks and scratch filters.

The programs pass from the studio controls, through the switching system, into one of three program channels, each having four inputs. Each channel input is provided with a listening key and mixer, while each channel has also a main attenuator. In addition, there are three channels of a simpler type having one input only and no attenuators. These three so-called "common channels" are used for rehearsals and for uncomplicated sound recordings requiring no special arrangements.

To stage, supervise, and control complicated broadcasts involving several studios, an arrangement called the "dramatic control" has been developed. It consists of a special control desk with six input channels and one output with the necessary controls. It may be connected between the studio control and the main control by the automatic switching system, and provides for full control of the broadcast. The dramatic control can handle programs from all studios, but is arranged to provide visual supervision of studios 6, 7, 8, and 9. This dramatic control also permits the introduction of variable equalizers to give special effects to certain studios or echoes. A studio amplifier similar to the *B* amplifier compensates for the losses introduced in the mixer.

3. Switching

All switching between program sources and channel or dramatic inputs is done automatically. The switching of microphone leads, volume controls, and signaling system is done in one operation, reducing substantially the possibilities for mistakes.

The switching system must have a very high degree of reliability, and at the same time must

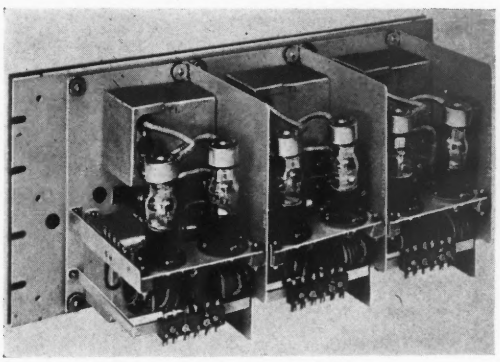


Fig. 3—A group of three *A* amplifiers mounted on a single panel. These amplifiers are connected to the microphones to provide both gain and isolation.

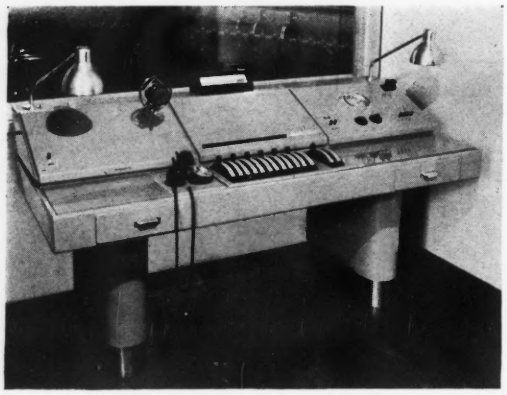


Fig. 4—Studio control desk.

be very flexible and simple in operation. Because of the amplification employed, switching must be entirely free from induced, capacitive, and microphonic noises. There are 21 channel inputs to which a maximum of 100 program sources may be connected. An entirely satisfactory system has been built with the same kind of apparatus and switches used in public and private automatic rotary telephone systems of Standard Electric manufacture. All relays are provided with split contact springs and double spherical contacts.

Switching is based on a coordinate system with relays at the intersecting points. The selection of the proper relay is determined by a control circuit that receives key-sent numbers and selects and operates the correct relay. The relay locks electrically and releases the control circuit for another call. To minimize the number of digits to be sent and to safeguard against mistakes, a channel of a particular control desk may be selected only from that desk. A distinction is also made between selection of a program and selection of an echo room; a three-position key in one position will permit only program selection and in another only echo-room selection.

Progress lamps on the control desks indicate when the control circuit is ready to receive the digits and when the selection is finished. Busy lamps indicate the channels in use. As an extra safeguard, each program channel or echo room is provided with a two-digit number-indicator operated from the control circuit during selection. It permits the operator to check immediately that the correct number has been sent and that the appropriate selection has taken place. The indicator remains in position until the connection is released, thus permitting the operator at any time to know which program source is connected to any particular channel. This relay connection is under the control of a locked release key on the main control desk, and the relay may be released by depressing this key momentarily.

Program connections may be set up prior to starting a broadcast by opening the attenuators of the channel, thereby permitting a change from one program to another by further operation of the attenuators only, without waiting for selection of circuits by relays.

In case of faults on lines, attenuators, or connecting relays, a program may be transferred from one channel to another on the same or another control desk without interruption. The new connection is made while the old is still in service, and the program faded from the faulty circuit to the new one.

Routine test circuits have been provided by which all the functions of the control circuits may be checked. Faults that might cause delay

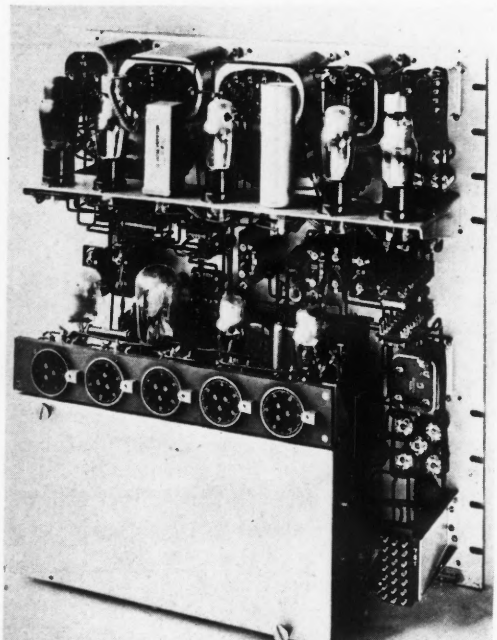


Fig. 5—The two-stage *C* amplifier, which also provides limiting action to prevent overmodulation. Two *C* amplifiers are mounted on a single panel.

in operation may thus be avoided. Two control circuits are provided and are utilized alternatively; thus, if one should be faulty, the second is always ready for use.

Particular care is taken to insure the system against cross talk and noise, by special screening and by keeping microphone and signaling leads completely separated throughout the system. No such troubles have been observed during the five years of operation.

Fig. 6—Main control desk. The amplifier room is visible to the operator through large triple windows.

A *C* amplifier is connected between the program source and the output switching panel. The *C* amplifier, shown in Fig. 5, consists of two two-stage resistance-coupled amplifiers, connected through a limiter to prevent the output level from rising above the value that corresponds to full modulation of the transmitter.

The program channels are operated from the three main control rooms. These are adjacent to the amplifier room, which may be viewed through large triple glass windows. Each of the three control rooms may operate one of the three program channels as well as all of the common channels. In each control room, all the different operations may be accomplished at the main control desk shown in Fig. 6.

For visual supervision of the program level, each channel output has been connected to a modulation meter with instruments on main and studio control desks. In principle, the modulation meter or tone meter is a rectifier obeying a logarithmic law and having a quick charge and slow discharge characteristic, so that peak voltages may be read. They are light-beam instruments with a very short time constant; being calibrated in nepers, the scale is practically linear. The upper part of the scale, corresponding to modulation of the transmitters above 80 percent, is red so that danger of overmodulation will be immediately apparent. When a studio is connected to a channel input, the instrument in the corresponding studio control room is automatically connected to the modulation meter of the channel in use.

Each control desk includes an output patching panel where, by means of a manual coordinate switchboard, the outputs of the *C* amplifiers associated with the channels are connected to the inputs of *D* amplifiers connected to the outgoing lines.

The *D* amplifiers are two-stage resistance-coupled amplifiers with low output impedance, so that they supply constant voltage to lines of varying impedances. Their outputs are connected to the "program consumers," i.e., the broadcasting transmitters, the Copenhagen repeater station for broadcasts to foreign countries, the recording rooms, and the loudspeaker channels.

The lines used for recording pass from the *D* amplifiers to the four recording studios through special recording amplifiers, which are power amplifiers similar to the loudspeaker amplifiers. Two of the recording studios are each provided with two wax recorders and two cellulose recorders, and have in addition play-back apparatus for cellulose and ordinary commercial records. The wax recorders are combined with play-back apparatus. The other two sound-recording studios are provided with tandem

magnetophone equipment to permit continuous magnetic recording and reproduction using plastic tape. A reproducer for commercial sound motion pictures is also available. A small announcer's studio is provided near the recording rooms to be used when recording or playing back programs.

Loudspeakers have been installed in studios, studio control rooms, and main control rooms for supervision of programs and for talk-back from studio control rooms to studios. The input to this system is taken from a separate *D* amplifier connected to each channel output. Feedback in the studios is avoided by attenuating the studio loudspeaker as soon as the attenuator of any microphone in the studio is opened. When talking back from control room to studio, a microphone placed on the control desk is cut in by a key, which also attenuates the control-room loudspeaker output.

A loudspeaker distribution system serves all offices, Council Hall, canteen, etc. The system accommodates four program channels, thus enabling the staff to listen to rehearsals as well as the running program. The loudspeakers require no amplifiers as the program lines are operated at high level from 40-watt amplifiers manufactured at the shops of the Post and Telegraph Administration.

Over the frequency range from 30 to 10,000 cycles per second, the total amplification from microphone to line does not deviate more than ± 0.2 neper (1.7 decibels) from the value at 800 cycles. At a total amplification in the system of 9.5 nepers (82.5 decibels), the noise voltage at the output of a *D* amplifier will not exceed 5 millivolts. In the range from 100 to 10,000 cycles, the distortion factor for the complete amplifier chain must not exceed 1 percent and from 30 to 100 cycles it must be less than 4 percent. The distortion factors are, in fact, considerably lower.

4. Signaling

A signaling system provides for the essential cooperation among all factors concerned with a broadcast transmission. The automatic switching system connects all the signaling and program circuits at the same time, thereby assuring readiness of the signaling system when required and avoiding separate operations with their possibilities of error.

The system includes red warning lamps and blue end-of-program lamps together with the necessary operating keys. The red lamps are displayed in the studio proper, on a signal stand in the studio, outside the studio entrance, on the studio control desk, and on the channels on the main control desk. The blue lamps are displayed on the studio signal stand, the studio control desk, and on the channels of the main control desk. Mounted on the wall in the three control rooms and in the sound recording rooms are signal boards on which the red and the blue lamps of each studio are represented.

An instant before a program is to go on the air, the control-desk operator—studio or main—in charge of this broadcast gives a microphone warning signal by flashing the red lights and then operating them steadily. In the case of delay in the studio, the studio manager may operate the microphone blocking key and continue the flashing, thereby indicating that the connection may not yet be made. When ready, he will restore his key and produce a steady red light. At the end of the program, the studio manager will switch on a blue light, the end-of-program signal, and the control desk will open the connection, thereby extinguishing the red light.

Connections among the various points of importance, such as studios and control desks may also be obtained by telephone through a private automatic branch exchange. As it may be of importance to obtain immediate communication with a certain party, the connection will be established even if the called line is busy; a warning tone will indicate that a third party is on the line.

The studios, buildings, and amplifier equipment were designed by the Broadcast Department of the Post and Telegraph Administration and manufactured by Standard Electric A/S, Copenhagen, who also designed and manufactured the switching equipment, desks, loudspeaker cabinets, and similar facilities.

Zurich Automatic Tandem Toll Exchange

By G. KLINGELFUSS

Standard Telephone and Radio S. A., Berne, Switzerland

DURING the past 20 years, very great progress has been made in the automatization of the telephone service in Switzerland. Now, about 95 percent of the 450,000 subscribers have automatic switching service, and within a few years the entire network will be automatic.

After the city and rural networks were converted to automatic operation, most of the fundamental problems of long-distance automatic telephony, such as impulse transmission over long lines and metering of calls by time and distance, were solved. The next advance was to provide automatic service between subscribers of different network groups.

As a temporary solution, long-distance dialing by operators was introduced, whereby the toll operator at the outgoing end dialed directly the number of the wanted subscriber in another city.

The first application of direct automatic toll switching by subscribers in different city areas equipped with rotary-system exchanges was introduced in 1933 between Basle and Zurich. Similar services were provided between Berne and Lausanne with Strowger equipment. The first application of interworking between two different systems was between Lausanne and Geneva with Strowger and rotary equipment, respectively.

The good results obtained on these automatic toll circuits encouraged their extension. A fundamental plan for automatic operation of the Swiss toll network had been prepared by the Swiss administration and was elaborated on jointly by Bell Telephone Manufacturing Company and Standard Telephones and Cables, Limited, giving full consideration to switching, signaling, and transmission requirements.

The Swiss telephone network includes three different automatic systems, Hasler, rotary, and Siemens. It was necessary, therefore, to establish fundamental requirements for their interworking. For this purpose, the Swiss administration prepared a master specification that contained all the requirements essential to reliable interoperation.

These specifications concerned not only signaling, but transmission requirements over long-distance lines, with fixed repeaters and with switched repeaters connected at tandem points. The repeater problems were still to be solved, particularly, switched repeaters with automatic gain control. The Swiss administration adopted for all three systems the principles of switched repeaters developed and recommended by Bell-Standard.

The three manufacturers were requested to submit models of every circuit involved in automatic toll switching. The three models were interconnected over real toll-line circuits of the greatest length met with in Switzerland using both fixed and switched repeaters. For signaling, 50-cycle-per-second alternating current was used.

The Swiss automatic toll network in its final stage is shown in Fig. 1. It is comprised of 10 primary centers and about 40 secondary or end exchanges. Each toll center is equipped with repeaters and associated switching apparatus. The toll end exchanges have direct trunks to the toll centers to which they are assigned.

Three-digit prefixes starting with 0 were assigned to the primary and secondary exchanges (Zurich 051). The third digit determines whether the call is destined to remain in the primary center group (51) or has to be extended to an end exchange assigned to this center (Winterthur 052, Schaffhausen 053).

1. Toll Center Zurich

Zurich is the largest and most important toll center of Switzerland. It handles about 20 to 25 percent of the total telephone traffic and consists of 11 city exchanges of 10,000 lines with 3 satellite exchanges of the 7-A1 and 7-A2 rotary system and 50 rural exchanges of the 7-D type. It may be interesting to note that the first rotary exchange at Zurich-Hottingen was put in service in 1917 and is still in operation.

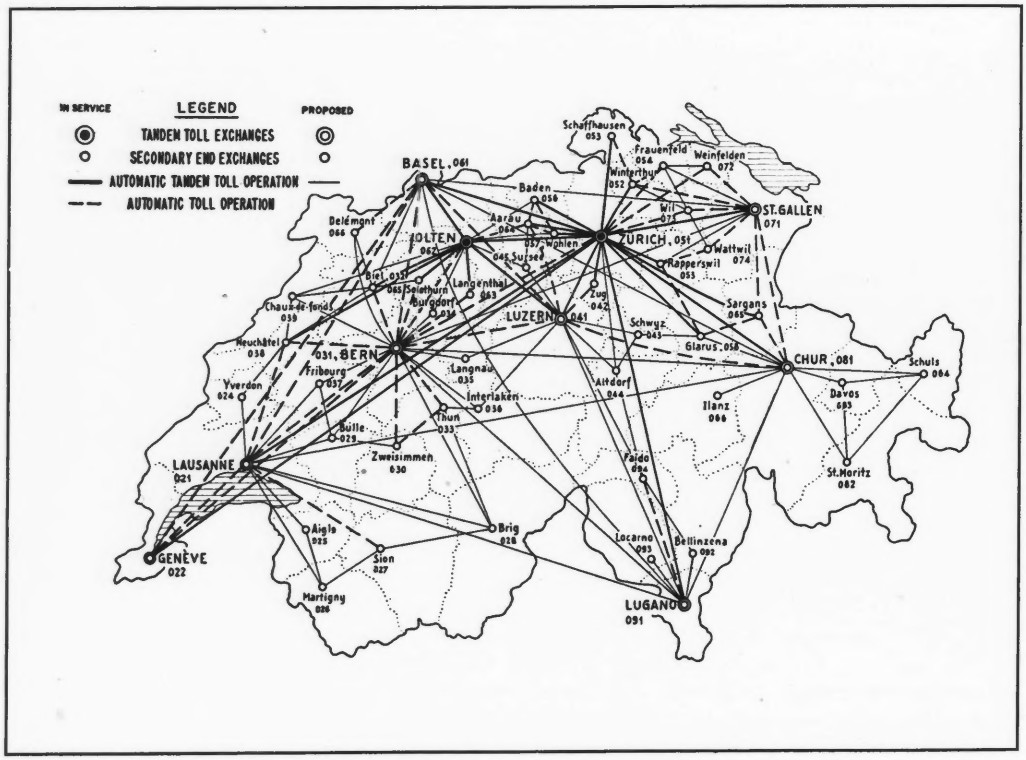


Fig. 1—Automatic toll network of Switzerland.

Zurich also has a very important national and international toll traffic, being the largest commercial and industrial center of the country. Consequently, it is also the most important toll switching point.

The first equipment for the present automatic toll exchange was ordered in 1937. A new building, shown in Fig. 2, was erected for this purpose next to an existing building, in which the manual toll exchange,

Fig. 2—Building housing the automatic toll exchange and repeater station in Zurich.

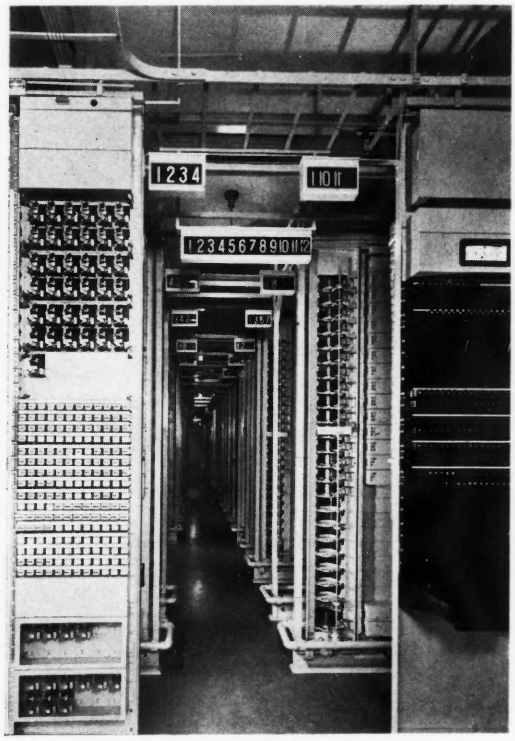


Fig. 3—General view of the automatic toll exchange.

rural main exchange, and repeater station are located.

The automatic toll equipment, shown in Fig. 3, is located on the third floor, and the repeater station is on the floor directly below. This arrangement requires only very short cable runs to the repeaters, which are located in both the automatic toll switching room and in the repeater station.

The first installation of automatic toll equipment, which was placed in service in 1939, consisted only of apparatus for the outgoing terminal toll traffic originated by Zurich city and rural subscribers and of equipment for handling connections established by manual toll operators on incoming national or international toll lines. The equipment for incoming terminal traffic destined for Zurich city and rural subscribers is located in the old building and was also extended and adapted to the administration's master specification.

During the war, several extensions were made and, in 1945, equipment for automatic tandem toll operation was put in service, whereby for the first time automatic switched repeaters were used in the rotary system. Through this development, Zurich became the largest and most important tandem point for automatic toll switching.

A simplified junction diagram of the Zurich automatic toll exchange in its present stage is shown in Fig. 4. The actual equipment occupies a floor space of 550 square meters (5918 square feet), and completely fills the room. Future extensions will have to be located on the floor above.

2. *Toll Traffic Originated by City and Rural Subscribers*

To establish an automatic toll connection, a subscriber first dials a three-digit toll prefix (021 to 099) followed by the number of the wanted subscriber, which may be of two, five, or six digits.

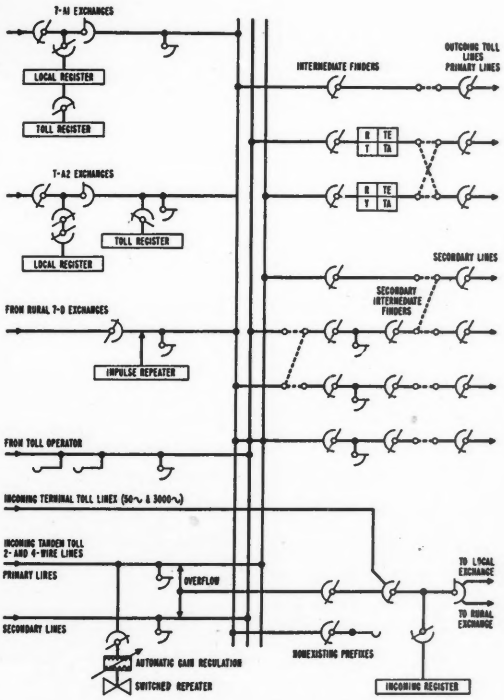


Fig. 4—Simplified junction diagram of the Zurich automatic toll exchange.

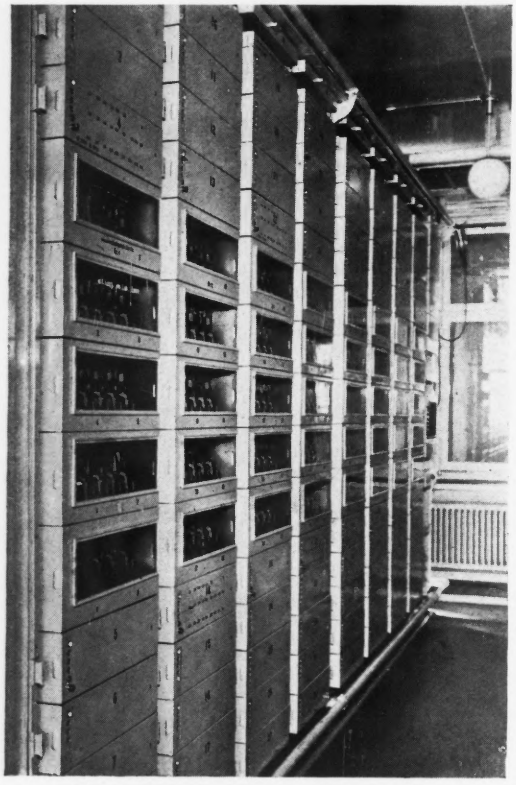


Fig. 5—Selective trunk circuits from 7-A1 local exchanges.

With the dialing of 0 as the first digit, the calling subscriber is extended through a selective trunk circuit to the automatic toll exchange.

3. Selective Trunk Circuits

There are three different classes of selective trunk circuits:

A. Toll traffic originated by city subscribers,
B. Toll traffic originated by rural subscribers,
C. Incoming calls on manual or international toll lines that are to be extended to the automatic toll network.

The selective trunk circuits from the city exchanges are divided into two different classes: incoming from 7-A1, shown in Fig. 5, and from 7-A2 exchanges.

In the 7-A1 exchanges, a toll or auxiliary register is attached to the local register when a subscriber dials 0 as the first digit. The auxiliary register remains attached until all digits are sent out.

For the selective trunks from 7-A2 exchanges, the toll register is located in the automatic toll exchange and the 7-A2 local registers are arranged for sending up to 9 digits into the toll register.

The selective trunks from the local exchanges are combined with a time-and-zone metering circuit and are equipped with three step-by-step selectors of which one serves for the selection of the direction, one for determining the tariff according to the dialed toll prefix, and one for producing a number of artificial tandem selections as may be required.

A tariff-control circuit is also attached to handle networks with mixed tariffs.

The selective trunk circuits for the automatic toll traffic originated by rural subscribers are inserted between the rural main exchange in Zurich and the automatic toll equipment. During the selection time, an impulse repeater is attached to the selective trunk circuit and is controlled by the auxiliary register in the rural center exchange. This type of selective trunk circuit has no time-and-zone metering circuit because the tariff is determined in the outgoing rural center exchange.

The selective trunk circuit for the toll traffic from the manual toll board is similar to that from the rural network. It appears on the toll-operator positions and also on twin jacks on the cord-circuit repeater positions for connections that require a cord-circuit repeater with automatic gain control. The toll operator dials only two-digit toll prefixes, because the 0 is not required in this case.

There are a number of common circuits used with the selective trunk circuits, such as, master time-control circuit, impulse sender circuit, and delayed back release circuit.

4. Incoming Toll Lines

The incoming toll lines are divided into two classes, which are fundamentally different: terminal lines and tandem lines.

The incoming terminal lines carry only traffic destined for city and rural subscribers of the Zurich group.

The incoming tandem lines are for the transit traffic passing through Zurich to other toll primary or toll end-exchange groups. Overflow

facilities are provided so that when all terminal lines are busy, traffic may enter over tandem lines.

The incoming tandem lines are further divided into primary and secondary groups. The primary or main lines carry the incoming traffic from other primary groups, whereas the secondary lines handle traffic from the toll end-exchange groups.

The tandem primary lines are again subdivided into 2-wire and 4-wire lines, which are shown in Figs. 6 and 7.

The 2-wire lines are arranged for 50-cycle signaling, whereas the 4-wire lines, which are used for operation on 12-channel carrier systems, are arranged for 3000-cycle signaling.

The incoming tandem toll lines have access by means of a finder to the switching equipment for repeaters with automatic gain control.

The 2-wire lines are equipped with individual networks.

Each incoming tandem toll line is provided with a switch for connecting the line in the wanted direction, as determined by the digits received from the far end.

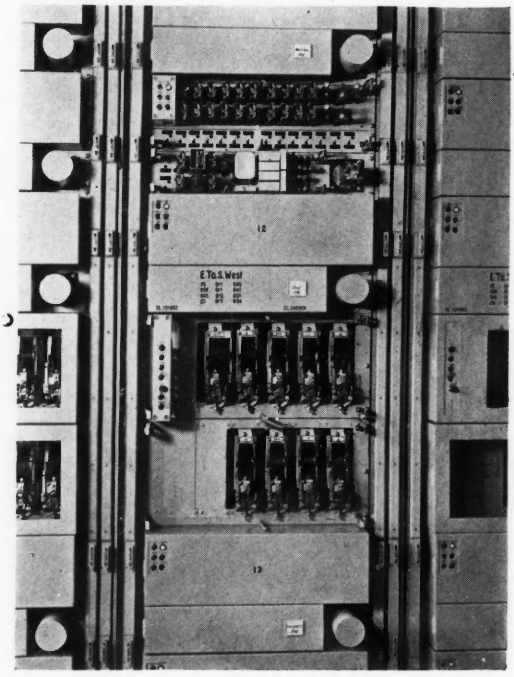


Fig. 6—Incoming 2-wire tandem toll lines.

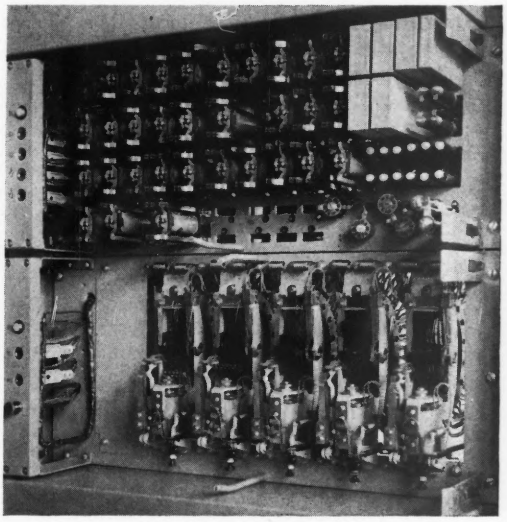


Fig. 7—Incoming 4-wire tandem toll lines.

The incoming tandem lines are connected to the arcs of the intermediate finders via cross-connecting points.

Incoming terminal traffic on tandem lines in case of overflow, is routed through another group of intermediate finders, incoming cord circuits, and incoming registers to the local or rural exchanges, as the case may be.

5. Intermediate Finder Circuits

The intermediate finder circuits, shown in Fig. 8, are used for concentrating the traffic between the selecting equipment and the outgoing toll lines. As will be seen from the junction diagram, there are three divisional groups of intermediate finders on the arcs of which are terminated:

A. Selective trunks,
B. Incoming tandem primary lines,
C. Incoming tandem secondary lines.

In the direction of secondary toll centers, there are two stages of intermediate finders. The first intermediate finders include equipment for the selection of the wanted direction. These may be seen in Fig. 9.

The intermediate finders assigned to the incoming tandem toll lines are arranged to provide

for overflow from outgoing terminal lines to outgoing tandem lines (*X* and *Y* directions) in case all terminal lines in a wanted direction are busy.

A further group of intermediate finder circuits are provided for routing incorrect toll calls for nonexistant toll prefixes to an operator's position.

The intermediate finders are arranged for gradual starting and, to reduce hunting time, starting occurs at the same time the finders of the free outgoing toll lines of the wanted direction are set in motion.

6. Outgoing Toll Line Circuits

Each outgoing toll line is terminated on a backward-hunting finder. These circuits, illustrated in Fig. 10, are subdivided into two main groups, viz., primary lines leading to other primary or toll centers and secondary lines connected to secondary or toll end exchanges.

The lines of the first group are again subdivided into terminal and tandem lines for each direction according to the transmission grade. Overflow facilities are provided and, when all terminal lines in a certain direction are busy, the terminating traffic will be routed over tandem lines of the proper direction. This operation is determined by the preceding selective trunking equipment.

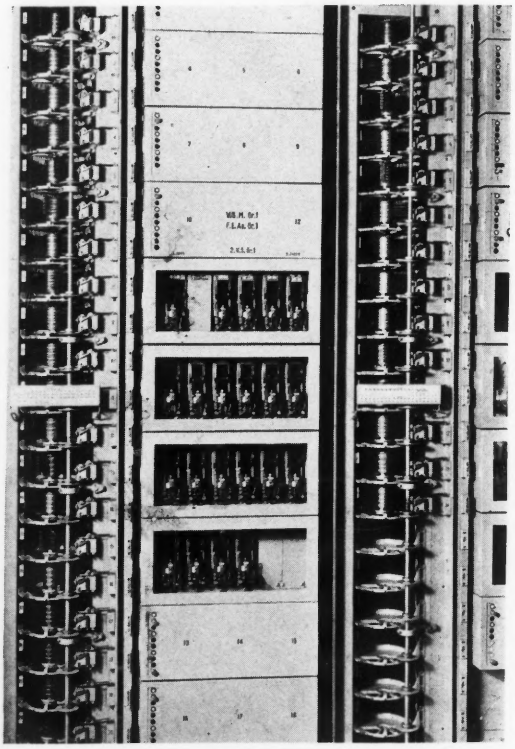


Fig. 9—Intermediate finder circuits with direction markers.

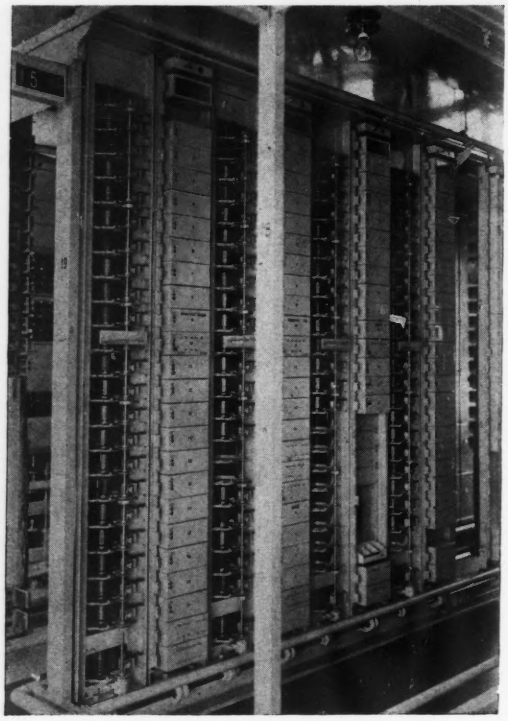


Fig. 8—Intermediate finder equipment.

On directions with a great number of lines, gradual starting circuits are provided so that only a limited number of free finders will hunt simultaneously when a certain direction is selected.

The outgoing terminal lines are grouped into *TE1* and *TE2* lines. The *TE1* terminal lines, to which only the outgoing traffic originated by the Zurich city and rural subscribers is routed, are not equipped with networks as such connections never require the insertion of a switched repeater.

The *TE2* terminal lines to which incoming tandem lines also have access are, however,

equipped with individual networks in case tandem connections with switched repeaters are to be established. For this purpose, each line circuit is also equipped for line-loss signaling.

The line-loss indication will always be signaled because the circuit does not know where the switched repeaters are connected. This signal is transmitted at the moment the outgoing toll line is connected, i.e., normally at the same time the starting impulse is sent towards the distant exchange.

Direct-current impulses received by the outgoing toll circuit from the local or rural exchanges are translated into alternating-current impulses for transmission to the distant end of the outgoing toll line.

7. *Switched Repeaters and Automatic Gain Control*

Switched repeaters are automatically introduced on tandem connections. Standard 2-wire repeaters have been adapted for this purpose. Each switched repeater circuit is combined with a gain-control circuit consisting of a number of relays and two variable pads for the incoming and outgoing lines. The finder, by which the automatic repeater is connected to the incoming line, consists of 11 brushes of which 2×4 serve for the through connection of the talking and network leads of the incoming and outgoing side.

The repeater connecting and regulating equipment is located in the switch room, whereas the repeaters and gain-control circuits are mounted in the repeater room on the floor below.

The automatic-gain-control circuits may be seen in Fig. 11 and each consists of four variable pads, which can be regulated in 1-decibel steps up to 15 decibels. The switched repeaters are set to give a gain of 17 decibels (1.96 nepers). In the idle position of the repeater, all pads are inserted so that the effective gain is reduced to 2 decibels. As soon as a repeater is switched in, the gain is automatically adjusted to meet the circuit needs.

For setting gain, three distinct indications are required.

A. Loss value of the incoming toll line,

B. Whether the gain of the switched repeater is to be adjusted to compensate for the loss in the incoming toll line or to equal the sum of the loss values of both toll lines, minus 9 decibels (1 neper),

C. Loss in the outgoing toll lines. When the gain-control relays receive a loss indication from the incoming toll line only, the gain of the repeater is set to equal this loss.

On the connections terminating in the Zurich local or rural area, no repeater is used. This is indicated by the setting of the marker switch of the incoming line.

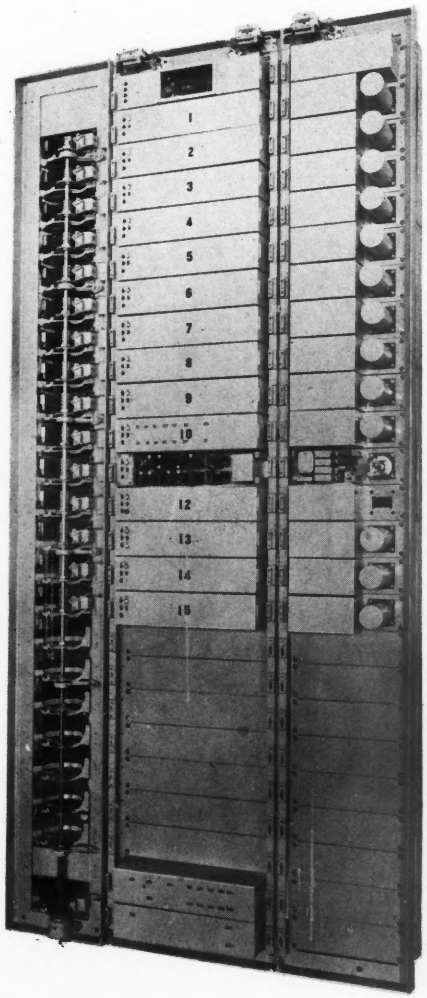


Fig. 10—Outgoing 2-wire toll lines with 50-cycle signaling sets.

8. Signaling Systems

Two different systems of signaling are used in the Swiss automatic toll system.

For relatively short 2-wire lines, 50-cycle signaling is used. The signaling current is relayed to bypass the line repeaters.

With the introduction of 12-channel carrier systems in the toll cable network, it became necessary to develop a new voice-frequency signaling system. For this purpose, Standard Telephone and Radio and the Swiss administration jointly have developed a single-frequency 3000-cycle signaling system.

9. Tandem Connections Over 4-Wire Lines

With the introduction of 12-channel carrier systems, incoming and outgoing 4-wire line circuits were provided to allow the following through connections:

A. 2-wire to 4-wire lines,
B. 4-wire to 2-wire lines,
C. 4-wire to 4-wire lines.

The incoming and outgoing 4-wire lines are terminated on 4-wire terminating sets.

When 4-wire lines are connected to 2-wire lines, the 3000-cycle signaling impulses incoming on the 4-wire circuit are retransmitted as 50-cycle impulses to the outgoing 2-wire line, and vice versa.

For connections between 4-wire lines, the tail-eating method is used and allows simple switching by avoiding the use of a great number of relay contacts.

10. Installed Equipment

The Zurich automatic toll exchange has the following number of principal circuits.

351 Incoming selective trunk circuits from 7-A1 local exchanges,
117 Incoming selective trunk circuits from 7-A2 local exchanges,
100 Incoming selective trunk circuits from 7-D rural exchanges,
20 Incoming selective trunk circuits from manual toll operators,
344 Incoming 2-wire tandem toll lines,
150 Incoming 4-wire tandem toll lines,
460 Incoming 2-wire terminal toll lines,
180 Repeater-connecting and gain-regulating circuits,
1790 Intermediate finder circuits,
432 Intermediate finder circuits with markers,
702 Outgoing 2-wire toll lines,
150 Outgoing 4-wire toll lines.

With the rapid increase of toll traffic in Switzerland, the Zurich automatic toll exchange will be developed further. International toll traffic will undoubtedly introduce new switching problems in automatic operation and these will be the subject of further studies in automatic toll switching.

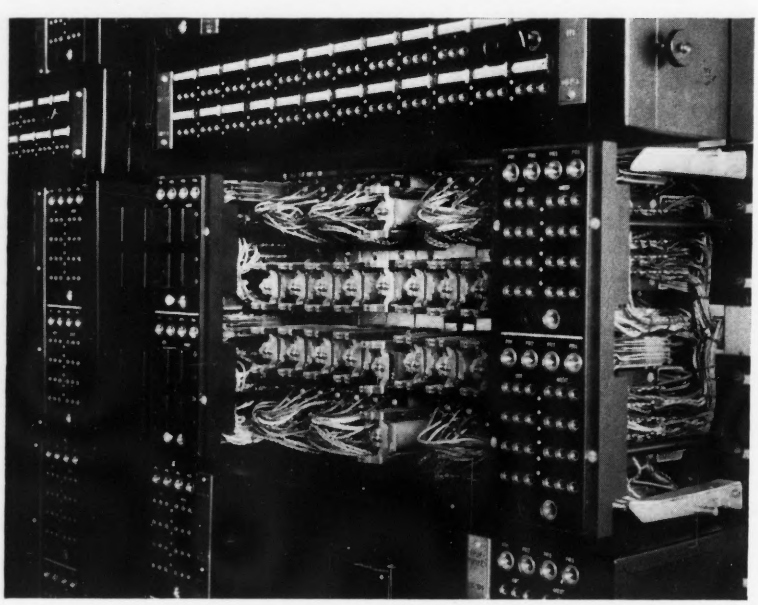


Fig. 11—Pads for automatic gain control of switched repeaters.

Emergency Operation of SOJ-12 Carrier Telephone Systems

By J. M. CLARA and E. SERRA CALABUIG

Compañía Telefónica Nacional de España, Madrid, Spain

PLANT EXTENSIONS for increasing the traffic capacity of the toll routes of Spain by the use of 12-channel open-wire carrier telephone systems were started in 1946. The carrier equipment is the *SOJ*-12 system manufactured by Standard Telephones and Cables, Limited, of London. A description is given of how the circuits on the Madrid–Barcelona route were restored on an emergency basis. This was facilitated by the self-aligning feature of the repeaters and receiving terminals. Although some increase in unintelligible intersystem cross talk would be expected under such conditions, this was not found to be serious enough to prevent commercial use of the circuits.

• • •

During the past ten years, well-known difficulties have delayed telephone development in Spain, where, commencing in 1924, the Compañía Telefónica Nacional de España, then an associate of the International Telephone and Telegraph Corporation, had achieved great progress in the development and quality of its service. Within the limitations imposed by world-wide difficulties, efforts were made to normalize the service and, before the close of World War II, plans had been formulated to provide quick relief for the congested toll service between the most important Spanish cities; this construction was to be started once the war was over.

The toll routes of Spain were totally reconstructed approximately twenty years ago; they have been kept at an extremely high grade of maintenance and, having repaired the damages that resulted from the civil war of our country, at the end of the war we had available a toll plant consisting mostly of open-wire line in first-class condition. It was therefore decided that the first major relief to be given for a period of 3 to 5 years would be based on applying single-, 3-, and 12-channel carrier systems to existing open wire, this having the additional advantage that it required the minimum use of copper, which was in very short supply.

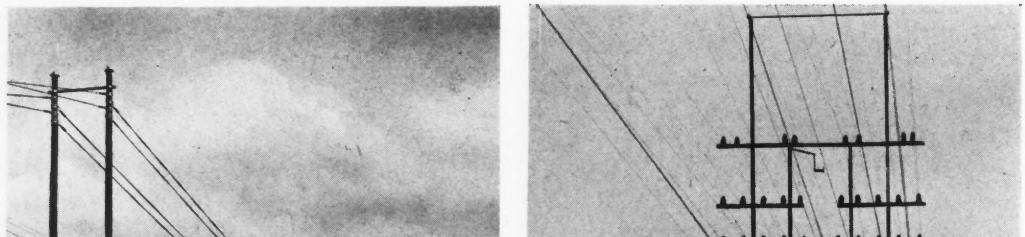
In 1943, a plan was developed for the installation of a number of *SOJ* 12-channel carrier systems, the installation of which is now well advanced. As a description of the over-all plan is beyond the scope of the present paper, we will limit the discussion to experiences during abnormal circumstances where the equipment worked successfully even though the operating margins for which it had been designed were considerably exceeded.

The first route on which *SOJ* 12-channel carrier was applied was Madrid–Barcelona, via Valencia, this having been modified to provide the necessary transpositions so that a number of systems could operate on the same pole line. Three *SOJ* systems were cut into service between October, 1946, and January, 1947.

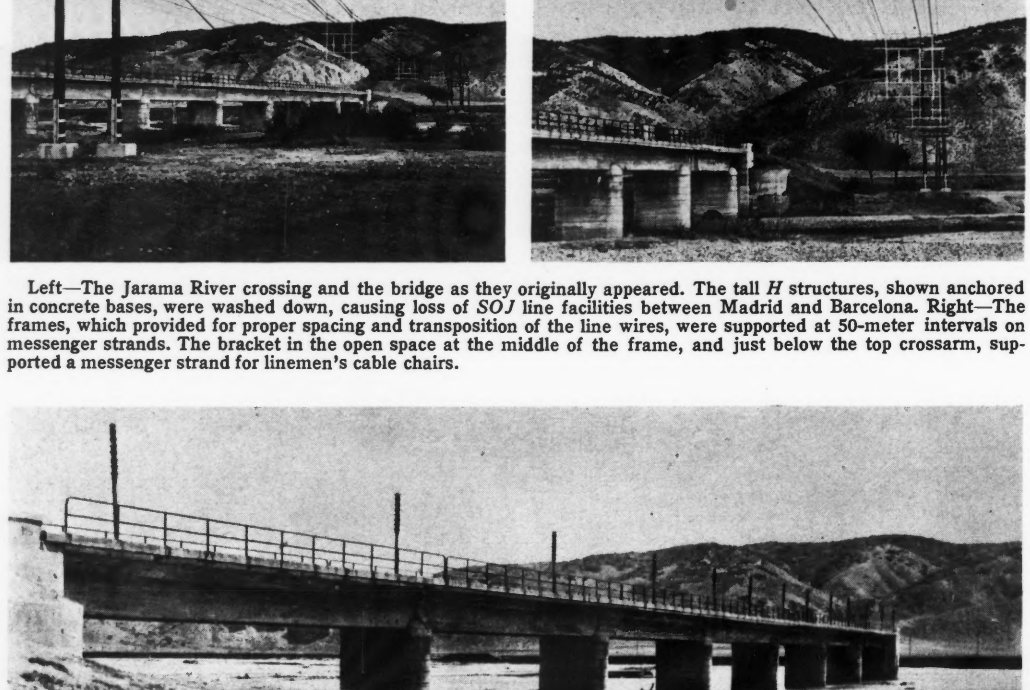
There existed some apprehension that in the event of a major line failure we would be in serious difficulties, since there were no alternate routes for 12-channel carrier operation. Unfortunately, this situation occurred in early March, 1947, when, as a result of exceptionally long-continued rains and melting snows, floods interrupted all types of communication in different parts of Spain.

At the point indicated in the accompanying map, the Madrid–Barcelona route crosses the Jarama River with a 200-meter (656-foot) open-wire span. To obtain correct wire separation and proper transposition spacing, the wires were supported by three crossarm fixtures suspended at 50-meter (164-foot) intervals from messenger strands fastened to special *H* structures at each side of the river. The *H* structures consisted of specially selected tall creosoted pine poles fastened to concrete bases located on the down-river side of the highway approaches to the bridge. In addition to wires, crossarms, and messenger strand, these poles supported a special strand to facilitate line maintenance over the river with the use of a lineman's cable chair.

The rising river flooded the valley and washed out some 200 meters (656 feet) of the eastern highway approach to the bridge, which remained



Left—The Jarama River crossing and the bridge as they originally appeared. The tall *H* structures, shown anchored in concrete bases, were washed down, causing loss of *SOJ* line facilities between Madrid and Barcelona. Right—The frames, which provided for proper spacing and transposition of the line wires, were supported at 50-meter intervals on messenger strands. The bracket in the open space at the middle of the frame, and just below the top crossarm, supported a messenger strand for linemen's cable chairs.



View of the bridge after the flood. Remains of one of the *H* structures are in the right foreground. All land between the end of the bridge and the distant right was washed out. The poles lashed to the bridge railing support the emergency lines.

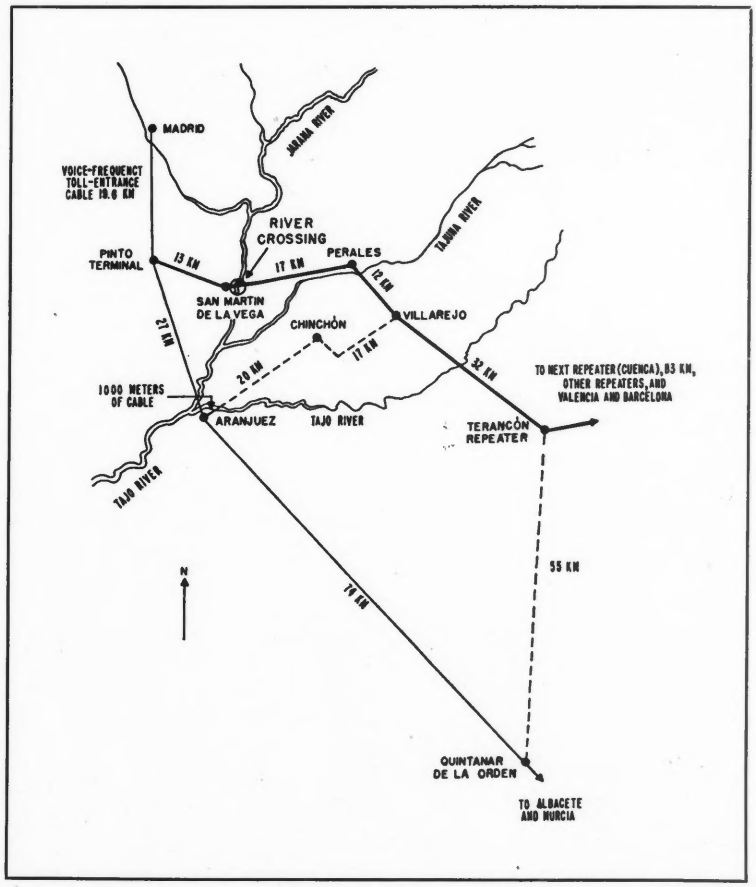
isolated. The force of the flood and the cutting of a new river channel caused the fall of the *H* structure at the eastern side of the bridge, and the line thus fell for a distance of nearly 600 meters (1969 feet).

The first indication of the river-crossing failure was observed at noon on March 5th, when the circuits on the lower arm and two pairs on the top arm were grounded. This put out of operation one of the three *SOJ* systems, and left no spare transposed pairs for carrier. As a trial emergency measure, the out-of-service *SOJ* system was transferred to a Pinto–Cuenca pair of the second arm, and the Tarancon repeater of this *SOJ* system was cut out of service; this latter being necessary because no patching arrangements existed at Tarancon to make use of the circuit for *SOJ*-system operation. In spite of the difference in transmission levels, and the fact that the pair was not transposed for carrier frequencies, the three *SOJ* systems functioned satisfactorily on this basis for seven hours until the rising river washed out the whole crossing.

Consideration was immediately given to the use of some other line facilities for the *SOJ* systems until such time as the normal route could be temporarily repaired. The only two possibilities were the Pinto–Aranjuez–Chinchon–Villarejo route (96 kilometers, or 59.7 miles), and the Pinto–Aranjuez–Quintanar de la Orden–Tarancon route (156 kilometers, or 96.9 miles), in which the Aranjuez–Chinchon–Villarejo and the Quintanar de la Orden–Tarancon sections are secondary or almost rural routes.

The emergency circuits were made up of 1000 meters (3281 feet) of 0.4-millimeter (0.16-inch) subscriber cable, and various sections of 2- and 3-millimeter (0.08- and 0.12-inch) copper open-wire line, partly on brackets.

Patching to obtain through connections was made with rubber-insulated twisted-pair drop wire of 0.8-millimeter (0.32-inch) phosphor-bronze conductor, this being readily at hand. Work was carried out with difficulty as the area was flooded and most highways and roads were cut. Nevertheless, at 7:00 PM the same day,

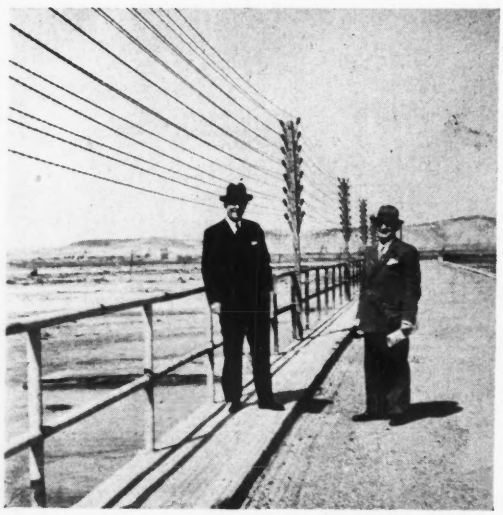


Map of the area in which emergency measures were necessary. The heavy line indicates the routing of the Madrid–Barcelona *SOJ* systems. The light line and dashed lines indicate a secondary or rural route, and the emergency interconnections, respectively.

an *SOJ* and a 3-channel system were placed in operation over the Pinto–Aranjuez–Quintanar de la Orden–Tarancon route, and at 3:00 the following morning, a second *SOJ* and a second 3-channel system were operating over the Pinto–Aranjuez–Villarejo–Tarancon route, thus giving 16 speech channels over each of the two emergency routes. No physical pairs were available for the re-establishment of the third *SOJ* system.

During these rerouting operations, preparations were made for temporary repair of the break at the Jarama River crossing on the normal main route. Light barges were hauled by truck to the nearest bridge approach, and in the early hours of March 6th an attempt was made to cross the river, but this was suspended owing to whirlpools and debris and the swiftness of the current.

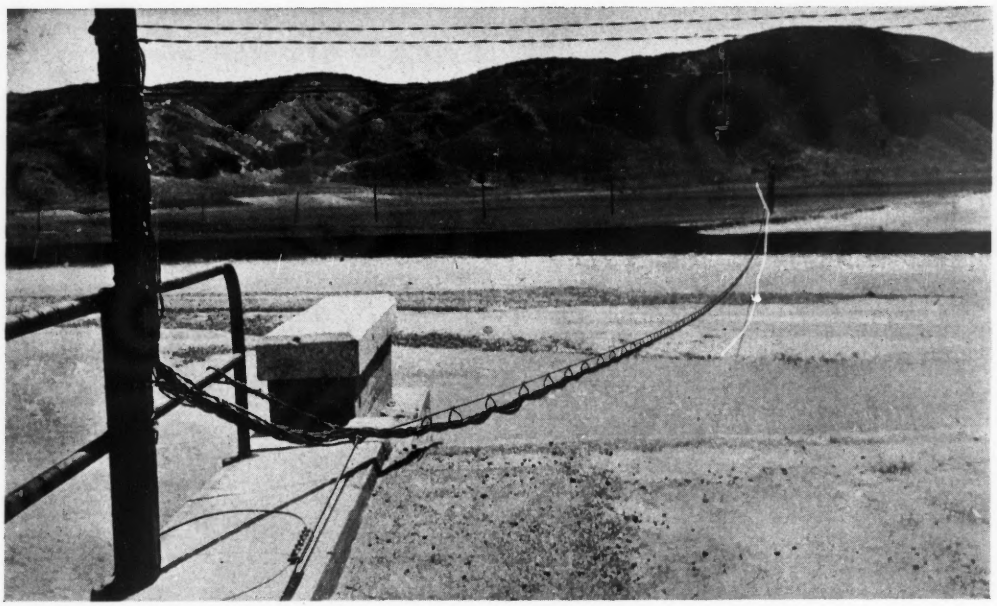
At considerable personal risk and discomfort, line wire, drop wire, and sky rockets were partly carried and partly swum to the Pinto side of the bridge, which had remained standing like an island in the river. From there a line attached to a rocket was shot over the new 200-meter-wide river channel to the Perales side. The rockets used were the largest that could be obtained of the type used in fireworks displays. Once this line was across, drop wires were pulled through this swift current and service re-established. It is interesting to note that with the passing of but one twisted pair across the river, 16 speech circuits (12 *SOJ*-system, one 3-channel-carrier, and one voice-frequency) were obtained. In fact, as the *SOJ* system had been previously turned on and was ready for operation, at the moment of connecting one wire of the river-crossing pair to open-wire temporary termination poles, all channels of the *SOJ* system commenced to operate even though the other wire of the circuit had not yet been spliced, so demonstrating one of the peculiarities of this high-frequency carrier equipment.



View of the emergency lines on crossarms with diagonal brackets. The authors are in the foreground.

The emergency wire used in this temporary repair consisted of rubber-covered twisted pair with 0.8-millimeter phosphor-bronze conductor. Approximately 500 meters (1640 feet) of this wire were spliced into each pair of the damaged section, half the length being fastened to the side of the bridge and the other half supported in a ring run suspended from a messenger cable strung over the new river channel; the river span was too wide for the drop wires to be self supporting. For the carrier pairs, one twisted pair was used for each conductor, these being separated from each other and the other wires by about 15 centimeters (6 inches) in the 250-meter (820-foot) section along the bridge. Separation between wires and circuits was, of course, impossible in the ring run over the new river channel. As each of these pairs were re-established, the three *SOJ* systems, as well as the other carrier systems, were transferred from their temporary rerouting back to this route; all functioned satisfactorily in spite of the impedance irregularity and close physical association of the wires at the crossing.

The permanent rebuilding of the Jarama River crossing will take considerable time, and once the river has completely subsided, a provisional reconstruction is proposed. However, at the time of this writing (May, 1947), after two months of operation over the emergency river crossing, the three *SOJ* and four 3-channel carrier systems continue functioning without any abnormalities, and with scarcely any impairment in quality of service. It will, of course, be necessary to reinstate the affected line section to its former standards of construction in order to ensure continued and satisfactory service and to maintain the previous standards of low inter-



Ring run over the new 200-meter channel. In spite of the loss of wire spacing in each circuit, and the close proximity of all circuits, the *SOJ* and 3-channel systems operated satisfactorily.

system cross talk under varying conditions of line attenuation with changes in weather.

Our former fears as to difficulties which might be encountered in the use of *SOJ* 12-channel systems in the event of a major line break not only have been dissipated, but we have also found that the characteristics of the 12-channel carrier equipment make it possible to re-establish a large number of communication channels over damaged line facilities with a speed difficult if not impossible to obtain where only voice-frequency equipment is available.

Nickel–Iron Alloy Dust Cores*

By S. E. BUCKLEY

Standard Telephones and Cables, Limited, London, England

THE DEVELOPMENT of dust cores to meet the increasing demand for use in telecommunications equipment is outlined. Requirements for such uses are defined, a high and constant permeability and low losses in the core being the most important. Amongst the materials used for dust cores described are electrolytic iron, Permalloy (nickel-iron), and molybdenum–Permalloy.

The constancy of permeability and its influence on design, the question of calculating the dust-core permeability from inherent properties of the metal, and core losses are discussed; some figures are given illustrating the result of reducing core-loss coefficients over the course of dust-core development.

The effective resistances at audio frequencies of dust-cored coils of various materials are compared.

• • •

1. Performance Requirements of Magnetic Dust Cores

The first large-scale uses of ferromagnetic materials in electrical engineering were principally in equipment for the generation and distribution of electrical power. In these applications, the use of magnetic material in sheet form in the construction of generators and transformers has been common practice for many years. With the continuous development of telephony and telegraphy, and related telecommunications systems, the scope of use of ferromagnetic materials was greatly increased. The performance required of magnetic materials in telecommunications equipment covers a very wide range and is often of a very specialized nature and the problems encountered are considerably more complex than those in power engineering. In this paper, the manner in which dust cores, more especially nickel–iron alloy dust cores, have been developed to meet some of these special requirements will be considered.

The general requirements for a magnetic dust core are as follows. The core permeability must be reasonably high and must remain constant over a wide range of induced flux density, and also over a wide range of frequency. The permeability must remain constant with time, and the variations due to temperature changes and to accidental overloads must be small. The total losses occurring in the core must be low, and in many instances it is important for the hysteresis loss to be especially small. The process of manufacture has to be controlled so that the various requirements are all met within narrow limits of tolerance. In addition to meeting the electrical performance required, the cores must be sufficiently robust to withstand the subsequent handling entailed in such operations as coil winding, adjusting, impregnating, and potting.

The manner in which the core performance requirements led to the development and commercial production of dust cores may be stated briefly. In order to reduce eddy-current losses in the core, subdivision of the magnetic material in a direction perpendicular to that of the magnetic flux is necessary. The practical achievement of this requirement is illustrated by core constructions in which the magnetic material takes the form of thin insulated laminations or fine wire. Although this construction results in a reduction of eddy-current losses it does not ensure constancy of permeability under the various conditions already mentioned. To obtain this constancy, the magnetic material must be subdivided to introduce air-gaps in a direction perpendicular to the plane of lamination; for example, in a ring core, radial air-gaps are necessary. Thus, for some applications, the core would need to be assembled from spaced sectors of material, the sectors consisting either of stacks of insulated thin laminations, or else of strands of insulated thin wire bonded together. This general type of construction is cumbersome and costly, although it was used at one time for

* Reprinted from "Symposium on Powder Metallurgy," The Iron and Steel Institute, London, England, pp. 59–63; May, 1947.

loading-coil cores for telephone lines. A more practical commercial solution to the problem was sought.

2. *Outline of Progress in Core Development*

It had been appreciated for a long time, as is shown by a suggestion made by Heaviside,[1] in 1887, that a core composed of finely divided iron insulated, for example, with wax might exhibit the desired properties. Although cores of this type can have low losses and other desirable characteristics, the permeability obtainable is very low. It was not until the advantages of using forming pressures of the order of 100 tons per square inch were demonstrated, that the large-scale production of high-permeability dust cores was achieved. This development by the Bell Telephone Laboratories[2] in America was given impetus during the first World War, when the supply from Europe of diamond dies for the drawing of fine iron wire was stopped, and hence the production of iron-wire cores was jeopardized.

3. *Electrolytic-Iron Dust Cores*

The outline of the core technique then developed was as follows. Pure iron was deposited electrolytically under conditions which yielded a brittle product. The brittle iron was ball-milled to produce iron powder, which was next coated successively with zinc and shellac, and then pressed at about 100 tons per square inch into cores of permeability 35. The cores thus obtained had losses which were low enough for the cores to be used in loading (Pupin) coils for insertion in telephone-cable circuits and in other high-quality coils included in telephone-circuit equipment. The type of core which had previously been used consisted of fine iron wire wound into toroidal form and then cut radially to introduce stabilizing air gaps. The performance of the dust cores was greatly superior to that of the wire cores.

4. *Permalloy Dust Cores*

After the publication, in 1923, of Arnold and Elmen's[3] paper on Permalloy, the magnetic properties of nickel–iron alloys attracted considerable attention. It had been demonstrated that after suitable treatment, including especial care in heat treatment, these alloys exhibited much higher permeabilities and yet a much lower hysteresis loss than previously known materials. Since the nickel–iron alloys in sheet form had such excellent properties, the possibility of obtaining improved dust cores by using the alloys in dust form was apparent. Various methods of producing alloy powders have been suggested. However, it is important to remember that the magnetic properties of a metal or alloy, in common with many other physical properties, are determined by the crystalline structure of the material as well as by its chemical composition. Thus, in many instances, a material which exhibits good magnetic properties when in the solid form may be reduced to powder but, owing to the treatment involved in the powdering process, the magnetic properties are completely changed. Hence the process must be developed to produce powder which retains as far as possible the inherent magnetic properties of the solid material. A number of processes were investigated for the production in powder form of nickel–iron alloys containing approximately 80 per cent of nickel. The first of these methods for large-scale production is described in a paper by W. J. Shackelton and I. G. Barber,[4] published in 1928. Nickel and iron are melted together and a controlled amount of sulphur is allowed to remain in the melt so that when the ingots are subsequently hot-rolled, and then allowed to cool, a thin layer of sulphide separates out at the grain boundaries. The relative proportions of manganese and sulphur in the material must be so arranged that the alloy is sufficiently malleable to roll while hot, and yet sufficiently brittle on cooling for slabs 1/4 inch in thickness to be readily broken up. Fig. 1 is a photomicrograph showing the polygonal crystals separated by the sulphide layer. On pulverizing the slab, disintegration occurs at the grain boundaries so that the particles of powder correspond generally to the crystals seen in the slab. Fig. 2 shows the appearance of the particles and confirms how they resemble the grains in the structure shown

[1] O. Heaviside, *The Electrician*, p. 302; February 11, 1887.
[2] B. Speed and G. W. Elmen, *Transactions of the American Institute of Electrical Engineers*, v. 40, p. 596; 1921.
[3] H. D. Arnold and G. W. Elmen, *Journal of the Franklin Institute*, v. 195, n. 5, p. 621; 1923.
[4] W. J. Shackelton and I. G. Barber, *Transactions of the American Institute of Electrical Engineers*, v. 47, p. 429; 1928.

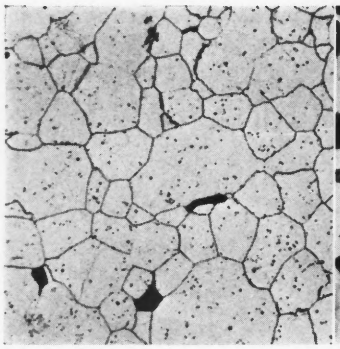


Fig. 1—Section of brittle molybdenum–Permalloy.

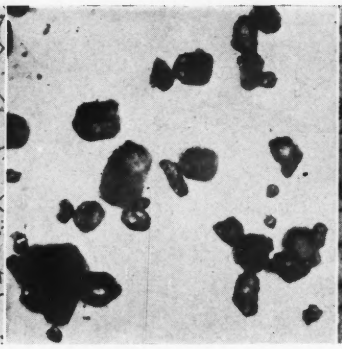


Fig. 2—Molybdenum–Permalloy dust.

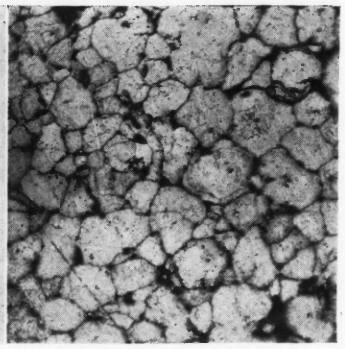


Fig. 3—Section of pressed molybdenum–Permalloy dust core.

in Fig. 1. Since the alloy has been work-hardened in the pulverizing process, it is annealed before proceeding to the stages of insulating and pressing. During the anneal some sintering occurs, but the original state of fineness is restored by light repulverizing. The annealed powder is next insulated by coating the particles with a thin layer of ceramic-type material which must withstand the subsequent pressing and core-annealing operations without rupturing. The insulated powder is pressed into cores at pressures up to 100 tons per square inch. The cohesion of these cores is derived mainly from the interlocking of the particles but, as already stated, the insulating layer must withstand this distortion during pressing. The high pressure which is necessary to produce close packing of the particles adversely affects the magnetic properties of the core, so that annealing of the pressed core is necessary in order to develop the desired properties. From powder containing approximately 80 per cent of nickel, dust cores of permeability 100 were produced. Although this permeability was much higher than that of the electrolytic-iron dust cores, the losses in the alloy dust cores were considerably lower, thus enabling a considerable reduction in core size to be made for a given coil performance.

5. Molybdenum–Permalloy Dust Cores

Further development of alloys of the Permalloy type in sheet form led to the development of nickel–iron alloys containing small percentages of other elements.[5] Thus 3.5 molybdenum–Permalloy, containing 3.5 per cent of molybdenum, has a higher initial permeability than the straight nickel–iron alloy, and has the additional advantages of lower hysteresis and lower eddy-current losses. Moreover, the heat treatment of molybdenum–Permalloy is much simpler and less critical than that of the straight alloy. The same general principles that were used for making 80-per-cent nickel–iron alloy dust cores were applied to making cores of 2:81 molybdenum–nickel–iron.[6] A number of improvements in the process were introduced and cores of permeability 125 were developed. Despite the higher permeability, the characteristics of these cores are superior in all respects to those of the previous type of permeability 100.

A section of a pressed dust core is shown in Fig. 3. Comparison with Figs. 1 and 2 shows the progressive stages by which the grains of the original solid material proceed through pulverizing to the final insulated and pressed state. The average particle size in the core is 0.002 inch.

6. Constancy of Permeability of Dust Cores

Fig. 4 compares the change of permeability with applied magnetizing force of sheet materials with corresponding dust cores. There has been considerable discussion as to the calculation of dust-core permeability from the inherent proper-

[5] G. W. Elmen, *Journal of the Franklin Institute*, v. 207, p. 583; 1929.

[6] V. E. Legg and F. J. Given, *Transactions of the American Institute of Electrical Engineers*, v. 59, p. 865; 1940.

ties of the material.[7] Although the matter is not yet finally settled, the following elementary treatment gives a good illustration of the effect, and claims to be no more than an illustration.

Let the relative length of magnetic particles to insulating material effective in the direction of the magnetic flux in the core be $l:a$. Let the permeability of the solid material be μ_1 and the permeability of the dust core be μ. Then, by the usual reluctance relationship for the magnetic circuit,

$$\frac{l+a}{\mu}=\frac{l}{\mu_1}+a.$$

From this,

$$\mu=\frac{\mu_1(l+a)}{l+a\mu_1}.$$

If the effective ratio of $l:a$ is 100:1, the following corresponding values of solid-material and dust-core permeabilities would result:

Solid Material:	1000	10,000	100,000
Dust Core:	92	100	101

[7] G. W. O. Howe, *Wireless Engineer*, v. 23, p. 291; November, 1946.

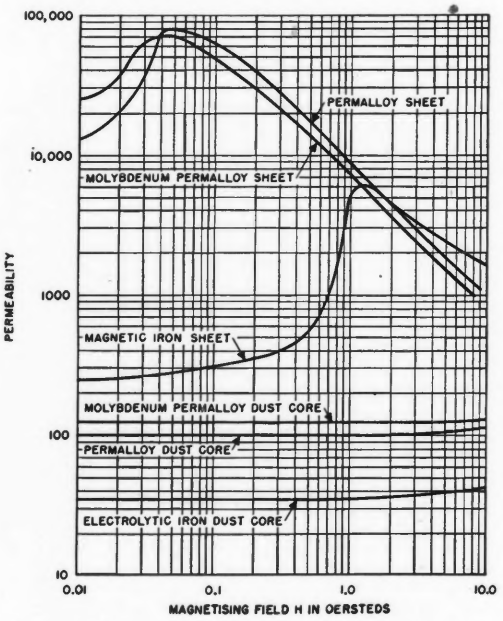


Fig. 4—Variation of permeability with magnetizing field for sheet and dust-core materials.

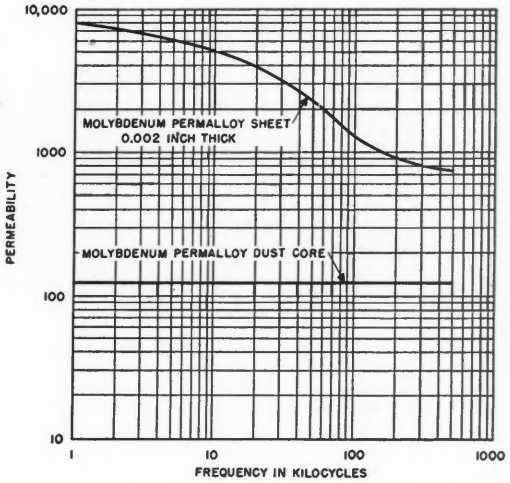


Fig. 5—Variation of permeability with frequency of thin molybdenum–Permalloy tape and dust core.

As already stated, the treatment given here is approximate only, and for further consideration the reference already quoted may be consulted, but the general manner in which the gap ratio has a stabilizing effect on the permeability of the dust core is demonstrated.

Similarly, Fig. 5, which compares the change with frequency of the permeability of thin tape material with that of a dust core, illustrates the stabilizing effect of the gaps.

7. *Core Losses of Dust Cores*

In electrical power engineering it is usual to express the magnetic core losses in the form of watts loss per pound or per kilogramme. In communications engineering, interest is usually centred on the effective resistance in ohms of a coil, so that an expression of core losses in the form ohms per henry (inductance) is generally used. It has been established[8,9] that the core loss of a dust core, or of a sheet magnetic material at low flux densities, may be expressed in the form

$$\frac{\Delta R}{\mu L}=(aB_m+c)f+ef^2,$$

[8] V. E. Legg, *Bell System Technical Journal*, v. 15, p. 39; 1936.

[9] H. Jordan, *Elektrische Nachrichten-Technik*, v. 1, p. 7; 1924.

where ΔR = core loss in ohms,
μ = core permeability,
L = inductance in henries,
f = frequency in cycles per second,
a = hysteresis coefficient of core loss,
c = residual coefficient of core loss,
e = eddy-current coefficient of core loss,
B_m = maximum flux density induced in the core.

The progress which has been made in reducing core-loss coefficients in the successive stages of dust-core development is shown in Table 1.

TABLE I

Material	Permeability	Hysteresis Loss, $a \times 10^6$	Residual Loss, $c \times 10^6$	Eddy-Current Loss, $e \times 10^9$
Electrolytic-iron dust	35	49	109	88
80:20 Permalloy dust	100	5.0	60	30
2:81 molybdenum–Permalloy dust	125	1.2	32	22

The reduction of hysteresis and residual losses is due to improvements in the magnetic quality of the materials. The reduction of eddy-current loss arises from higher electrical resistivity of the alloys.

8. *Resistance Characteristics of Dust-Cored Coils*

Fig. 6 shows the relative values of effective resistance at audio frequencies of coils wound on equal-sized cores (30 cubic centimetres) of the different materials. The great improvement obtained by using Permalloy cores is very apparent. It would be possible to design a coil on an electrolytic-iron dust core to have a direct-current resistance equal to that of the molybdenum–Permalloy dust-cored coil, but this would require the iron-dust core to be almost ten times as large. However, apart from direct-current resistance there are other considerations of very great importance. For many coil designs it is essential to restrict the hysteresis loss to a very low value. For equal hysteresis loss, the volumes of an iron-dust core and a molybdenum–Permalloy dust core would be in the ratio of 30:1. In practice, this means that in some applications it is now possible to use a dust-cored coil where previously no reasonable design was available. The

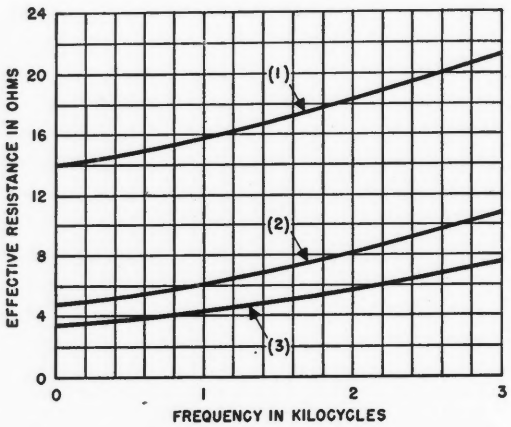


Fig. 6—Comparison of effective resistance at audio frequencies of dust-cored coils. All cores are 30 cubic centimetres in volume; coil inductance = 100 millihenries. (1), Electrolytic iron dust, permeability = 35; (2), 80:20 Permalloy dust, permeability = 100; (3), 2:81:17 molybdenum-Permalloy, permeability = 125.

usual procedure in replacing coil designs in the improved materials has been to produce improved coil performance combined with reduced size.

9. *Influence of Core Permeability on Coil Design*

Fig. 7 illustrates the resistance characteristics of coils wound on 30-cubic-centimetre cores of molybdenum–Permalloy of two permeabilities, 125 and 50. It will be noted that although the core losses of the 50-permeability core are ap-

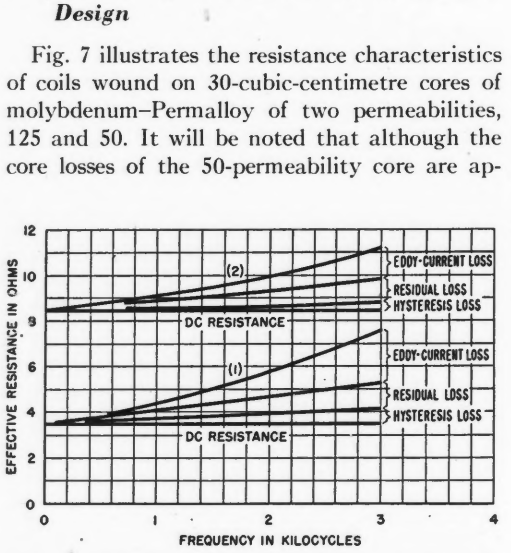


Fig. 7—Effective resistance at audio frequencies of molybdenum–Permalloy dust-cored coils. All cores 30 cubic centimetres in volume; coil inductance = 100 millihenries. (1), core permeability = 125; (2), core permeability = 50.

preciably lower than those of the 125-permeability core, the effective resistance of the former coil is considerably higher over the frequency range shown. The higher value is mainly due to the higher direct-current resistance. In order to reduce this resistance to that of the other coil, the core size would have to be increased by four times. This example shows the great importance of high-permeability cores of low loss in obtaining low-resistance coils of small size for audio-frequency use.

It is apparent from Fig. 7 that, although the two curves do not cross below 3 kilocycles, they will cross at a higher frequency above which the lower-permeability coil will have the better characteristics. Thus, when it is required to produce coils for use at various frequencies, it becomes necessary to produce cores of more than one permeability in order to cover the frequency range adequately. A measure of coil quality, that is often used and that enables the comparison of core materials to be shown conveniently on one graph for a wide frequency range, is the Q factor. The Q of a coil is the quotient of the coil reactance divided by the effective resistance, thus, $Q=\omega L/R$, and a high value of Q denotes a good coil. Fig. 8 shows the variation of Q with frequency for coils wound on 30-cubic-centimetre cores of various permeabilities. The magnetic material used throughout was molybdenum–Permalloy, and the general principles of coil manufacture are the same, the various permeabilities being obtained by different degrees of dilution of the core with nonmagnetic material. It will be seen that the 125-permeability core is best in the range up to 5 kilocycles, the 50-permeability core for the range 5–15 kilocycles, the 25-permeability core for 15–50 kilocycles, and the 13-permeability core from 50 kilocycles upwards.

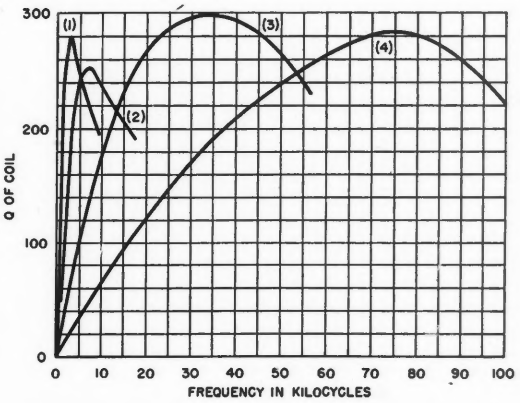


Fig. 8—Q values of coils on molybdenum–Permalloy dust cores of various permeabilities. All cores 30 cubic centimetres in volume; coil inductance range = 10–20 millihenries. Core permeabilities: (1) = 125, (2) = 50, (3) = 25, and (4) = 13.

10. Summary

This paper deals essentially with the properties of nickel–iron alloy dust cores made by the methods described. It is appreciated that there are other methods[10] by which alloy dust cores may be produced, and some information on these methods is already published. There is also a considerable range of low-permeability magnetic-dust cores produced by methods based on the general principles of plastic moulding. Information on these cores, which are widely used at radio frequencies, is also available from publications.

11. Acknowledgment

Thanks are due to Mr. E. L. Bodycombe for the photographs shown in Figs. 1, 2, and 3.

[10] G. R. Polgreen, *Post Office Electrical Engineers' Journal*, v. 37, p. 1; April, 1944.

Very-High-Frequency Single-Channel Receiver

By W. C. LANE and T. C. CLARK

Federal Telephone and Radio Corporation, Newark, New Jersey

INCREASED traffic at the major airports and the greater coverage required of airline enroute communications have shown the need for an improved very-high-frequency crystal-controlled ground-station receiver with high rejection of unwanted signals and good sensitivity. The receiver designed for these requirements operates on a single pretuned frequency in the range of 118 to 136 megacycles per second, and employs a 14-tube superheterodyne circuit, incorporating automatic-volume-control, squelch, and noise-limiter circuits. Low cost, small size, high quality, performance, and an easily serviceable unit were the points stressed in this design. Field tests indicate the effectiveness of the design.

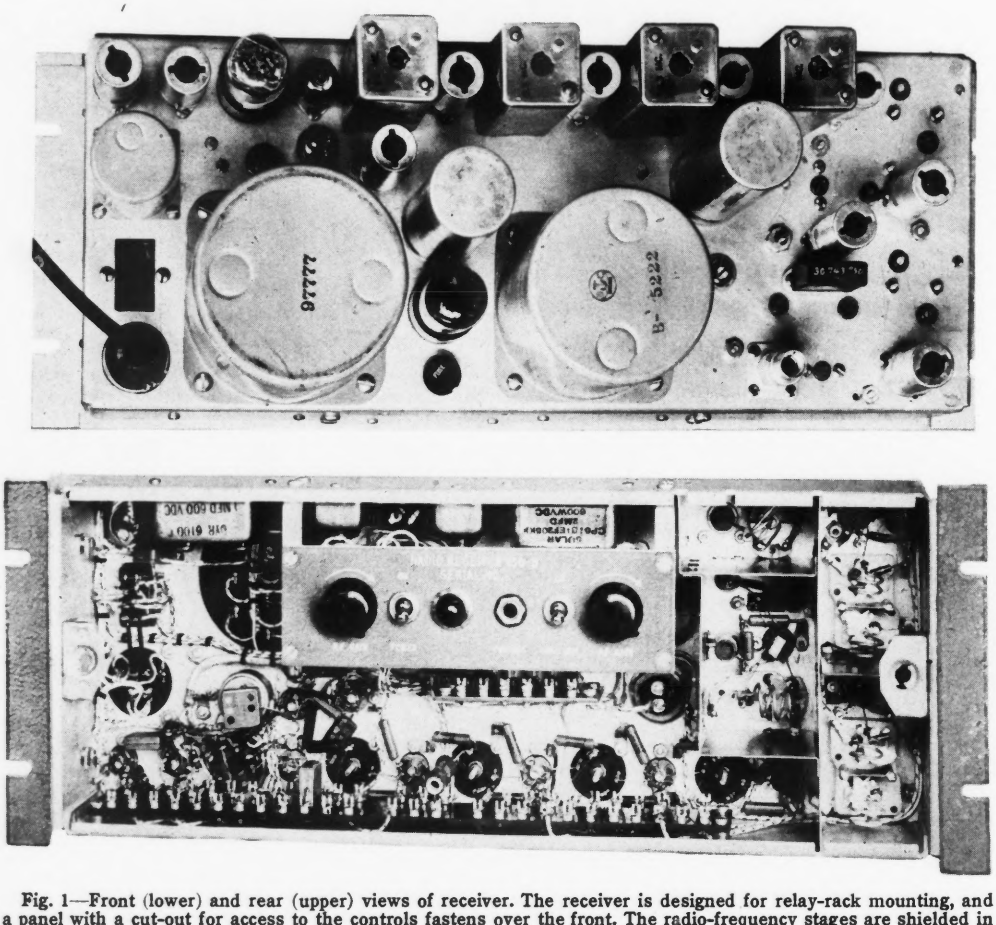


Fig. 1—Front (lower) and rear (upper) views of receiver. The receiver is designed for relay-rack mounting, and a panel with a cut-out for access to the controls fastens over the front. The radio-frequency stages are shielded in compartments (at the right).

1. Physical Characteristics

Mechanically, the receiver will mount in any standard relay rack and, since tubes and transformers are mounted horizontally at the rear, tube replacements and all tuning adjustments

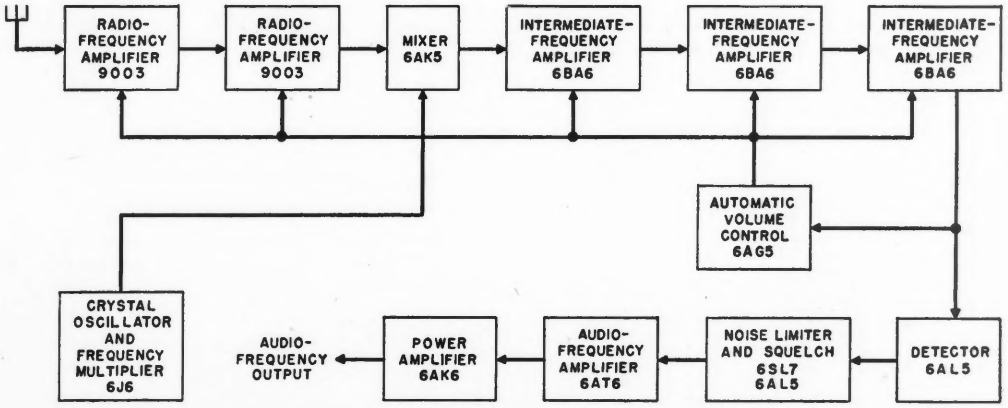


Fig. 2—Block diagram of single-channel 118–136-megacycle receiver. Squelch and noise-limiter stages improve the usefulness at airports, where there is generally considerable electrical noise.

can be made with a minimum of effort. The front panel is removable (Fig. 1) to facilitate inspection and repair without taking the receiver from the rack. Operating controls are mounted on a subpanel immediately behind a cutout in the front panel, and include an audio-frequency gain control, radio-frequency gain control, power on-off switch, squelch on-off switch, pilot lamp, and phone jack. Fig. 1 shows the unit-type construction employed.

2. Electrical Characteristics

Every effort was made in the design to use conventional and straightforward circuits in achieving the high-performance characteristics required. Fig. 2 is a block diagram of the receiver. The sensitivity of the prototype model was 0.7 microvolt for 50-milliwatts output at a 10-decibel signal-to-noise ratio. The image response was found to be down 86 decibels at the high-frequency end of the band, and the spurious responses resulting from oscillator harmonics were down 94 decibels or more.

2.1 Radio-Frequency Stages

Because of the relatively low intermediate frequency, which was necessary to achieve the required selectivity, it was found desirable to incorporate two double-tuned radio-frequency stages and a tuned antenna stage to attain high image rejection. The coils are constructed of 14-American-Wire-Gauge silver-plated copper wire, and are self-supported on the variable air capacitors for the shortest lead lengths and to obtain the highest Q. The 9003-type tube is chosen for radio-frequency amplifier, partly because of its remote-cut-off characteristic, which makes it readily adaptable to automatic-volume-control application, and also because of the small input loading. As an indication of the importance of input loading at these frequencies, two 9003's and a 6AK5 (used as amplifiers and mixer, respectively) were used to replace two 6BA6's and a 6AG5. With the tubes biased to the same transconductance and with the same coupling coefficient, the gain then increased 6 decibels and the image rejection improved 15 decibels.

Small baffle shields isolating the plate and grid pins at the tube sockets and a high degree of shielding between stages are necessary to minimize regeneration, and this allows an overall radio-frequency gain of 140 to be attained.

The bias on the radio-frequency tube grids is kept at the lowest permissible value to obtain the optimum signal-to-noise ratio.

frequency noise currents that were amplified by the sensitive receiver to produce objectionable noise output.

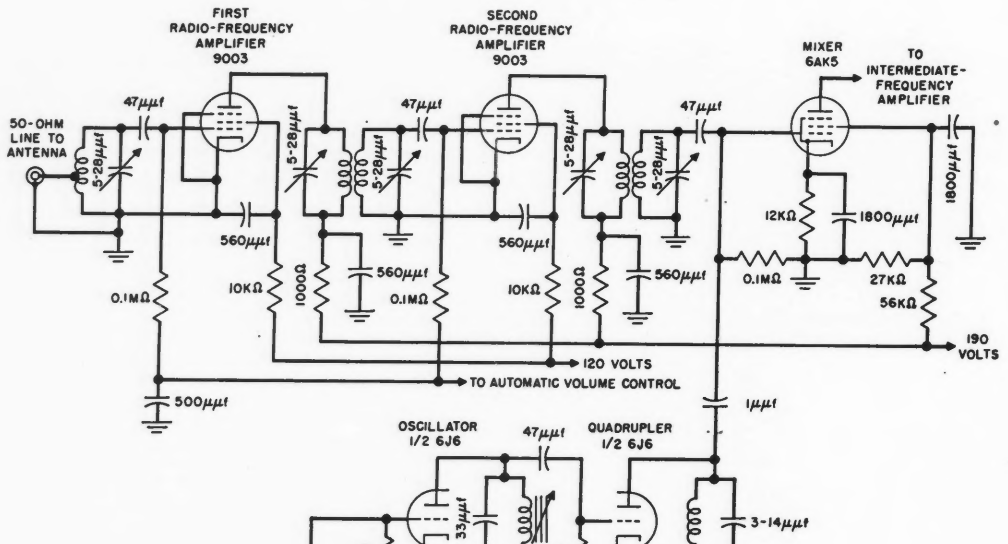


Fig. 3—Schematic of the radio-frequency stages and local oscillator. Congestion of the radio spectrum at airport locations requires good image rejection, sensitivity, stability, and selectivity.

2.2 Oscillator, Multiplier, and Mixer

The oscillator circuit (Fig. 3) is of the tuned-plate tuned-grid type; insertion of the proper crystal in the grid circuit will permit oscillation between 27.9 and 32.4 megacycles. The third-harmonic-type crystal is mounted in an unheated *CR-1* holder and has an over-all frequency tolerance of ±0.005 percent between temperature limits of −25 and +75 degrees centigrade (−13 and +167, Fahrenheit).

A third-mode crystal, with a built-in thermal unit to keep the crystal at a constant temperature, was employed on a previous model. This did not prove successful, since the opening and closing of the heater contacts set up radio-

A twin-triode tube, type 6J6, is used for the oscillator and multiplier. The crystal frequency is multiplied four times in a single stage, and is applied to the mixer grid through a 1-micro-microfarad coupling capacitor. Because of the high fundamental crystal frequency involved, the undesired third and fifth harmonics generated in the oscillator and quadrupler, and which are present at the mixer grid by an amount dependent on the quadrupler plate-circuit selectivity and the crystal activity, are far removed in frequency from the desired fourth harmonic. Thus, with the aid of the front-end selectivity, the spurious responses due to oscillator harmonics were practically eliminated. The oscillator funda-

mental and harmonic radiation is kept to a low value by proper shielding of the oscillator and quadrupler stages, and by filtering the power-supply leads to the compartment with button-mica feed-through capacitors. The radiation from the prototype model, as measured across the 50-ohm input circuit, was strongest at a frequency of 129.675 megacycles, for the channel frequency of 136 megacycles, and this was only 100 micromicrowatts.

The 6AK5 mixer operates with only 50 volts on the screen grid so that the injection voltage from the quadrupler may be small, but sufficient conversion gain may be obtained. The percentage of noise introduced by the mixer tube is negligible due to the high amplification preceding it.

2.3 Intermediate-Frequency Amplifier

The intermediate-frequency amplifier design was dictated by several considerations of selectivity: Present frequency allocations assign channels 200 kilocycles apart, so with the allowable combined drift of the transmitter and receiver, the bandwidth at the 6-decibel attenuation points must not be less than 48 kilocycles. In the event of 100-kilocycle channeling, the selectivity curve should afford at least 60 decibels of discrimination against an adjacent channel, which requires that the bandwidth at 60 decibels attenuation be no more than 180 kilocycles. By incorporating four double-tuned transformers in three 6.325-megacycle stages, the selectivity curve shown in Fig. 4 is obtained. A shape factor (which is the ratio of the bandwidth at 60 decibels to the bandwidth at 6 decibels) of 2.88 is indicated. Three remote-cut-off 6BA6 tubes are used in this system.

With the limitations imposed by present and possible future channel assignments, it is important that the intermediate-frequency system have a greater frequency stability than the preselector with regard to shifts of center frequency and bandwidth with changes in temperature, bias, line voltage, etc. In the design of this amplifier, iron-dust-core transformers with fixed silver-mica capacitors have been incorporated, and a high degree of temperature stability has been obtained. The center-frequency shift of the 6.325-megacycle intermediate-frequency system with a change in temperature from 0 to 40 degrees centigrade (+32 to +104 degrees Fahrenheit) is found to be less than 5 kilocycles. Due to the large tuned-circuit capacitance of 250 micromicrofarads, the Miller effect with gain-control bias variations is minimized and the center frequency shifts less than 4 kilocycles with a 60-decibel change in gain and a variation of ±10 percent in line voltage.

2.4 Automatic-Volume-Control System

The output of the automatic-volume-control amplifier is designed to regulate the grid bias of the two radio- and three intermediate-frequency stages in order that the output of the receiver will remain essentially constant with inputs between 2 microvolts and 0.1 volt (Fig. 5). Below the threshold value, the automatic-volume-control amplifier is cut off, so that

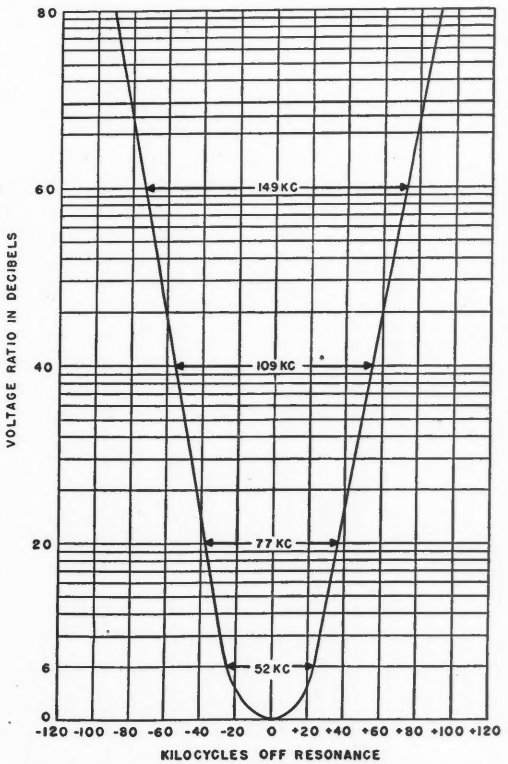


Fig. 4—Intermediate-frequency selectivity characteristic. Possible future assignment of 100-kilocycle channels in the band requires a steep slope at the sides of the curve. Center frequency is 6.325 megacycles.

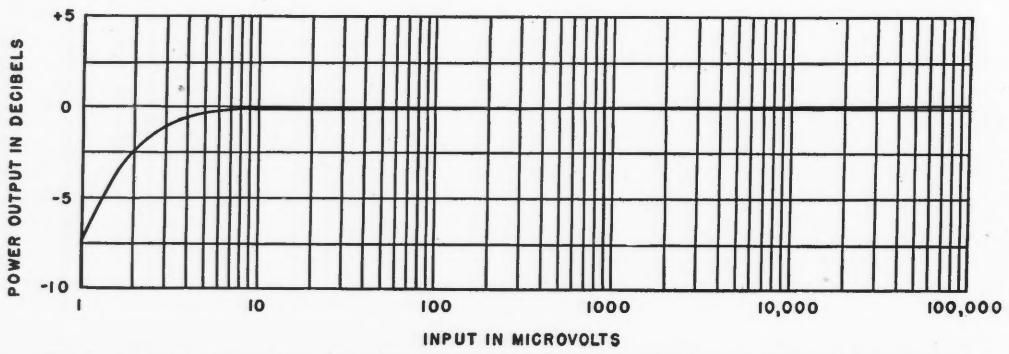


Fig. 5—Automatic-volume-control characteristic. Because of the generally low power of aircraft transmitters, and their widely variable distance from the receiver, the design must prevent loss of weak signals and blocking on the stronger ones. Zero level is 6 milliwatts in 600 ohms.

maximum receiver sensitivity prevails below the 2-microvolt level. This characteristic permits relatively weak signals to deliver a zero-level audio-frequency signal into a telephone line for remote operation. This feature is intended to preclude any need for remote radio-frequency gain control, although provision for that control is incorporated to provide for a remote squelch control.

2.5 Noise Limiter and Squelch

The noise-limiter circuit employed in this receiver is a very effective means of suppressing ignition noise, static, and other forms of recurrent pulses without distorting the audio-frequency signal. Fig. 6 shows the circuit, including the squelch, in schematic form. The theory of operation is explained as follows.

Under steady carrier conditions and with the squelch switch in the off position, the plate voltage of *V1* is more positive than that of *V2*, and therefore diodes *V3* and *V4* conduct. The relatively low impedance of diode *V4* allows the audio-frequency voltage to pass to the amplifier. Upon the application of a sharp noise pulse, the plate voltage of *V2* increases, but the plate voltage on *V1* remains momentarily constant due to the time constant of *R5* and *C2* in the grid circuit of *V1*. When the plate voltage of *V2* exceeds the plate voltage of *V1*, diodes *V3* and *V4* block, and the audio-frequency voltage is cut off for the duration of the pulse. An accelerating action, which tends to speed the blocking of the diodes, results from *C1* and *R6*.

The rectified noise pulse develops a negative voltage at the plate of *V4* and a more positive voltage on the cathode of *V3*, so that a fast blocking is obtained. When this occurs, the audio-frequency voltage at that instant will be held constant across *C3*, as there is no path for discharge. When the diodes again conduct, due to the decay of the noise pulse, the voltage across *C3* will then follow the instantaneous level. This provides a continuous audio-frequency signal with very little noise.

Voltage divider *R2* permits adjustment of the clipping level for modulation peaks from 60 to 100 percent. Although this control is normally set to pass a 100-percent-modulated wave, it is sometimes desirable in excessively noisy locations to improve the effectiveness of the noise limiter by reducing the bias on the grid of *V1*. This will tend to introduce distortion on highly modulated signals, but an improved signal-to-noise ratio will result and may be the more desirable.

By the addition of resistor *R10* in the plate circuit of *V1*, an inherent squelch action is obtained; with no signal input, the plate voltage of *V1* is lower than that of *V2* (because of the different values of plate-load resistance), and the diodes do not conduct. Upon arrival of a carrier of predetermined strength, dependent upon the amount of gain, the bias on *V1* becomes higher than the bias on *V2*, causing more voltage to appear at the plate of *V4* than at the cathode of *V3*, and the diodes conduct.

The radio-frequency gain control incorporated in the receiver allows control of the squelch for

signal inputs ranging from 1.5 to 1000 microvolts. This is an advantage to control-tower operators who want a muted receiver at noisy sites when no signal is on their frequency; they can adjust the gain control to bring a given noise level to the threshold of the squelch and still be confident of receiving weak signals.

2.6 Output Circuit

The audio-frequency amplifier system is capable of delivering a maximum of one watt into a 600-ohm output impedance. The audio-frequency fidelity is designed for speech operation and is flat to within ±3 decibels from 350 to 3500 cycles per second. The output transformer has a split-and-balanced secondary to provide maximum flexibility of load connections.

By simplexing the grid bias of the five gain-controlled tubes on the output circuit, the radio-frequency gain and squelch may be controlled remotely over a telephone line. *Simplexing* is a term meaning a system whereby a single 2-wire physical circuit may be made to carry independent direct-current and alternating-current signals; separation of these components being accomplished at the terminals by proper arrangement of inductances and capacitances.

A monitoring phone jack is connected across one section of the output-transformer secondary in order that the monitoring may be accomplished at the local installation independently of all external connections.

3. *Field Test*

The receiver has been installed and operated successfully in many control towers of major airports. In the Chicago airport control tower, sensitivity checks were made with Civil Aeronautics Administration cooperation. An outgoing plane flying at 4000 feet called the tower every five minutes. Due to heavy construction work in progress at the airport, a 5-microvolt noise level was induced in the antenna. The noise limiter incorporated in the receiver was so effective that signals from the plane's 5-watt transmitter were

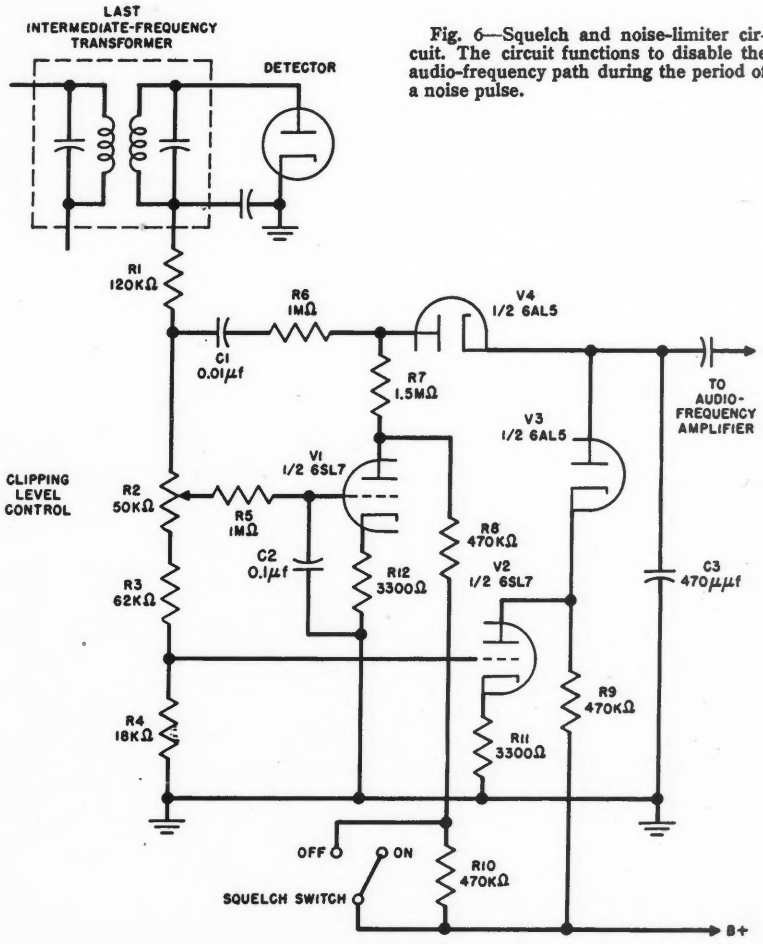


Fig. 6—Squelch and noise-limiter circuit. The circuit functions to disable the audio-frequency path during the period of a noise pulse.

received from a distance of more than 120 miles with a good signal-to-noise ratio.

In order to determine, under practical operating conditions, the effectiveness of the receiver selectivity in rejecting signals 100 kilocycles off resonance, field tests were arranged with the cooperation of American Airlines Incorporated. An antenna installed on the Lincoln Building (750 feet high) in Manhattan, was connected by about 50 feet of Type RG-8/U cable to the receiver, which was tuned to 126.7 megacycles. Carriers received from LaGuardia Airport (8 miles away), where a 50-watt transmitter was operating on 126.7 megacycles, developed a signal of approximately 1200 microvolts at the receiver input terminals. An identical receiver, adjusted for 126.8 megacycles, was then substituted. Due to the high order of adjacent-channel attenuation, no signal was heard on this receiver even though the squelch circuit was rendered inoperative and the receiver gain was at maximum.

Using the same receiving-antenna installation, the test was repeated under more severe conditions by using a 50-watt transmitter (126.7 megacycles) installed in the Lincoln Building. The spacing between the transmitting and receiving antennas was about 100 feet. This transmitter developed over 0.2 volt at the receiver input terminals. No blocking of the 126.7-megacycle receiver was noted. When the 126.8-megacycle receiver was substituted, an output signal equivalent to that produced by a 70-microvolt signal at resonance was obtained. Hence, the adjacent-channel selectivity attenuated the 126.7-megacycle 0.2-volt signal by more than 70 decibels. These test results therefore confirmed the laboratory measurements, and demonstrated the effectiveness of the radio- and intermediate-frequency amplifier design.

High Gain With Discone Antennas*

By A. G. KANDOIAN, W. SICHAK, and R. A. FELSENHELD

Federal Telecommunication Laboratories, Incorporated, Nutley, New Jersey

EXTENSION of work on discone-type antennas, originally presented before the Institute of Radio Engineers in 1945, is described. Since that time, these antennas have been developed for a large number of applications. Of particular interest are those applications that require high gain over a very broad band of frequencies. Detailed data are given on several designs, particularly on unidirectional antennas for point-to-point communication and omnidirectional high-gain antennas for the distance-measuring equipment (DME) of the Navar series of aids to aerial navigation.

• • •

At frequencies above about 30 megacycles per second, it is possible to produce, conveniently and economically, a great deal of directivity of radiation. For some services, high-gain antennas must also provide omnidirectional radiation in the horizontal plane. The amount of antenna power gain obtainable is then limited because concentration of only the vertical radiation is permissable.

Such antennas have been developed and widely used in very-high-frequency broadcasting and in radar beacons in the ultra- and super-high-frequency spectra. As a rule, however, the bandwidth requirements have been relatively small. For most applications, horizontal polarization has been used.

1. Omnidirectional High-Gain Broad-Band Antenna

The present antenna development was undertaken for the ground station of the Navar distance-measuring equipment (DME). Among the requirements specified were a frequency range of 960 to 1215 megacycles, vertical polarization, omnidirectional horizontal radiation pattern, power gain of approximately 8 over a half-wave dipole, and a beam direction tilted up approximately one-half beam width with respect to the horizon over the complete frequency range. There were also mechanical and service requirements common to such applications.

The basic radiating unit selected to meet these requirements is the discone antenna,[1] which in itself is an omnidirectional vertically polarized radiator with broad-band characteristics well in excess of the present requirements. By stacking a number of discones, one above the other, and supplying them with equal power in proper phase, the vertical pattern may be compressed and, hence, the over-all antenna gain greatly increased. An array of this type of nine discones is shown in Fig. 1.

The radiation pattern of such an array is given by

$$F(\beta) \cong \cos\beta \frac{\sin n\left(\frac{S^\circ}{2}\sin\beta - \frac{\phi}{2}\right)}{\sin\left(\frac{S^\circ}{2}\sin\beta - \frac{\phi}{2}\right)}, \qquad (1)$$

where β = vertical angle, zero at the horizon,
n = number of discones,
S° = spacing between successive radiators in electrical degrees, and
ϕ = phase difference between successive radiators.

When all the radiators are operated in phase, ϕ is zero and $F(\beta)$ is maximum for $\beta = 0$; that is, the radiation is always maximum in the horizontal direction. Fig. 2 shows a calculated pattern of such an array of 9 stacked elements. The power gain of an array of stacked discones is essentially the same as for an array of stacked horizontal loop antennas as given in reference 1.

Energizing a number of radiators in phase can be accomplished in many ways for a fixed frequency or a limited frequency band. Over a large frequency range, this may only be accomplished by making all the transmission lines to the individual radiators of equal length. This makes the relative phase of the currents in the various radiators equal and independent of frequency.

* Presented, National Electronics Conference, Chicago, Illinois, November 4, 1947. Reprinted from *Proceedings of the National Electronics Conference*, v. 3, pp. 336–346; 1947: figures 12, 13, and 14 were not included in the original publication.

[1] A. G. Kandoian, "Three New Antenna Types and Their Application," *Proceedings of the I.R.E.*, v. 34, pp. 70W–75W; February, 1946: Also, *Electrical Communication*, v. 23, pp. 27–34; March, 1946.

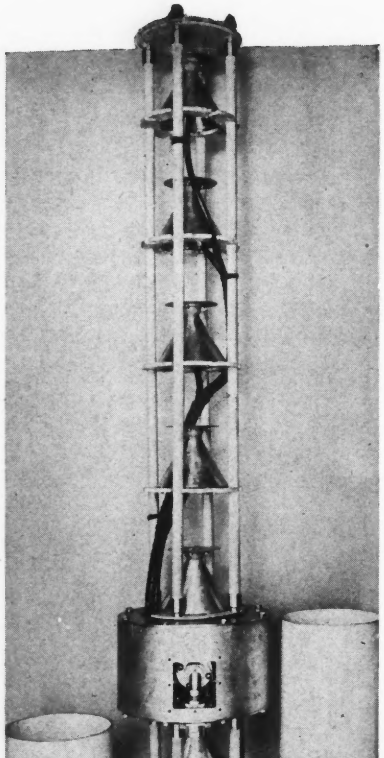


Fig. 1—Discone array used for the distance-indicator function of the Navar system. Consisting of 9 discone radiators stacked vertically, this omnidirectional antenna operates anywhere within the range of 960 to 1215 megacycles. The mountings and supporting rods are constructed of fibre-glass, as are the covers standing at the sides of the array.

In the present case, it is desired to tilt the pattern approximately one-half beam width above the horizon throughout the frequency range. This may be accomplished by supplying

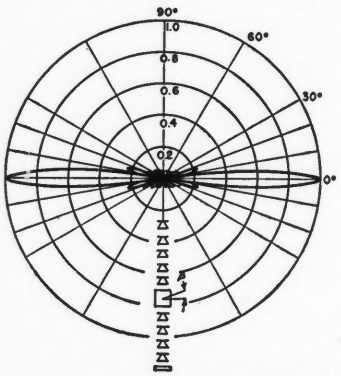


Fig. 2—Calculated vertical radiation pattern of a stack of 9 discones all energized in phase;

$$F(\beta) \leqq \cos\beta \frac{\sin\left[n(140\sin\beta)\right]}{\sin\left[140\sin\beta\right]}.$$

equal power to all the radiators as before, but successively delaying the phase of the current in each element from the bottom to the top. This relative phase is represented by ϕ in (1). The value of ϕ for a given tilt of pattern is

$$\phi = S^\circ \sin\beta_T, \qquad (2)$$

where β_T is the tilt of the beam with respect to the plane perpendicular to the antenna axis. This is derived simply by maximizing (1).

Thus, to produce a tilt in the beam direction of 3.5 degrees and for spacing S° equal to 280 degrees between successive discones, a phase stagger of $\phi = 15$ degrees is obtained. Fig. 3 shows the calculated pattern of such an array of 9 discones.

This establishes the basic theory of a high-gain omnidirectional antenna whose beam direction may be shifted electrically.

To reduce this to practice, first let us consider the simpler problem of an omnidirectional broadside antenna without beam tilting. The major problem is to divide the available power equally among a large number of elements and keep the currents in phase over a wide frequency range. This is accomplished in the present case by making the input impedance of each radiating element equal to the surge impedance of some

standard RG-type cable. The discones used in this array meet this requirement very well in that when used with RG-8/U cable they have a mismatch of less than 1.2 over a frequency range

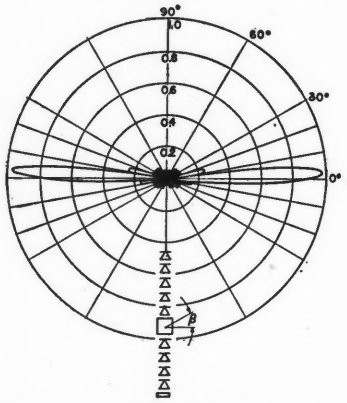


Fig. 3—Calculated vertical radiation pattern for array of Fig. 1 operating at 1087 megacycles;

$$F(\beta) \cong \cos\beta \frac{\sin n\left(\frac{S^\circ}{2}\sin\beta - \frac{\phi}{2}\right)}{\sin\left(\frac{S^\circ}{2}\sin\beta - \frac{\phi}{2}\right)}, \quad \text{and} \quad \phi = S^\circ \sin\beta_T.$$

much larger than 960 to 1215 megacycles. The individual feeders are all connected in parallel at a common junction box and then transformed back to 50 ohms impedance by a simple quarter-wave transformer. The bandwidth limitation of the array is now determined by the quarter-wave transformer, rather than the radiators themselves and, wherever justified, larger bandwidths could be obtained by multiple sections or exponential transformers.

The next step, that of tilting the beam above the horizon, is now very straightforward. All it is necessary to do is to make successive lengths of RG-8/U cable leading from the common junction box to the individual discones such that each one is 15 degrees longer than the one below it. This amounts to a length differential of only three-tenths of an inch between successive feeders.

There are certain possible complications in an array of this type, which tend to alter the input impedance and the radiation pattern. The major problem is the placement of the individual feeders. It was determined experimentally that least

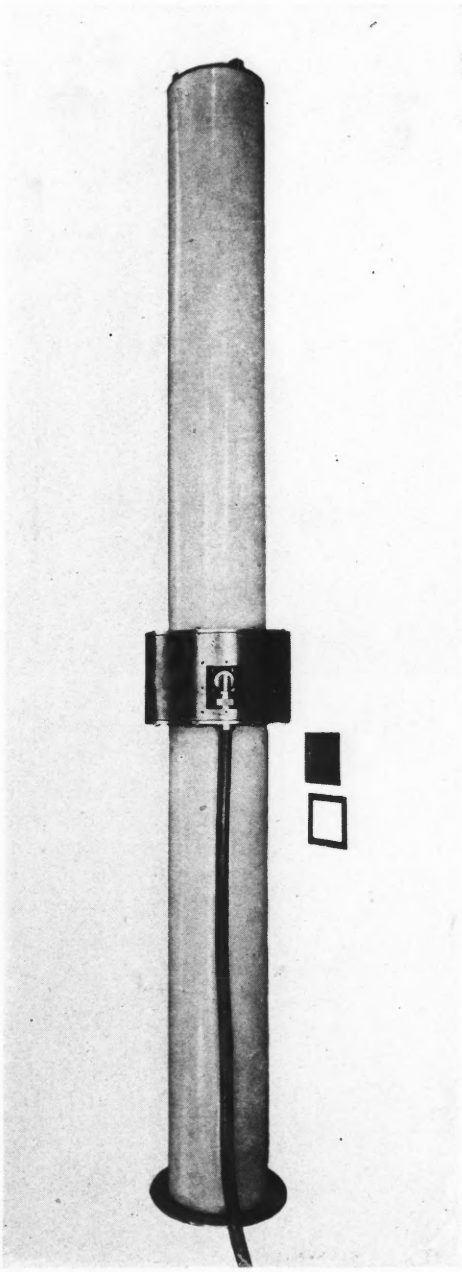


Fig. 4—Complete 9-element discone array with weather-proofing covers mounted. The access plate has been removed from the cable-connector housing, showing the type-N connector.

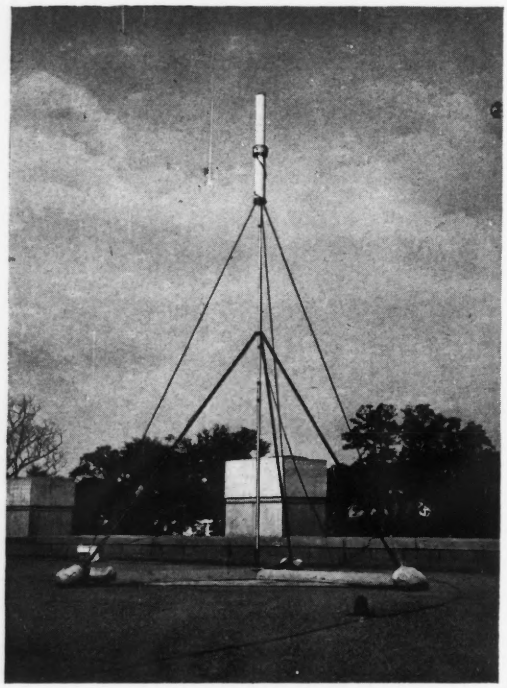


Fig. 5—Array mounted on mast on laboratory roof for flight tests with distance-measuring equipment.

disturbance was noticed when the cables were strapped together and then wound in a spiral around the discones, starting from the center of the array and working toward the top and bottom. The bunching gives the least reflection and the spiral tends to minimize the pattern distortion by spreading the effects evenly around the antenna. The excess length of cable feeding the center radiators is coiled up inside the metallic housing near the center of the array.

Mechanically, the antenna array is mounted on a fibre-glass supporting structure and enclosed from the weather by a fibre-glass housing. This results in a very strong and electrically efficient design. Fig. 4 shows the antenna unit complete in its housing. The over-all height of the structure is approximately 72 inches. Fig. 5 shows the array, 50 feet above ground level, installed in actual operation on top of a mast.

Measured data on the antenna array throughout the frequency range are given in Figs. 6 to 10. It is to be noted that though the half-power beam width narrows from 8 to 6 degrees from 960 to 1215 megacycles, the beam direction remains fixed at approximately 3.5 degrees above the horizon. The measured power gain was in all cases within ± 1 decibel of the calculated[1] value.

2. *Unidirectional Pencil-Beam Antenna*

In considering the problem of producing a unidirectional pencil-beam antenna as free as possible of any frequency limitations, the combining of a paraboloid and a discone antenna is a logical step. Measurements have shown that such a combination produces beam patterns that are satisfactory for communication and other applications over an extremely wide fre-

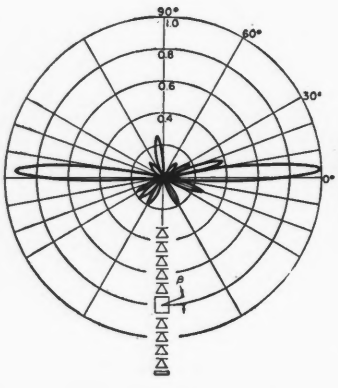


Fig. 6—Measured vertical radiation pattern at 960 megacycles.

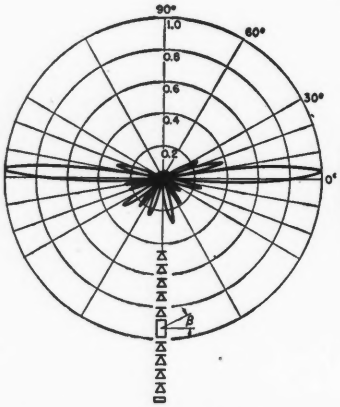


Fig. 7—Measured vertical radiation pattern at 1087 megacycles.

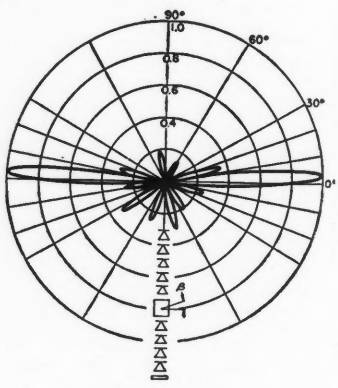


Fig. 8—Measured vertical radiation pattern at 1215 megacycles.

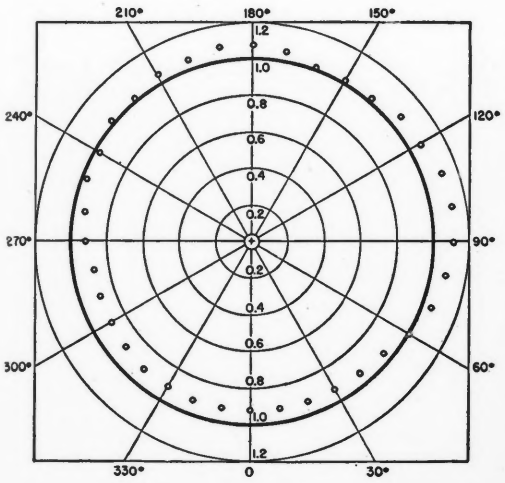


Fig. 9—Horizontal radiation pattern at 1087 megacycles. The calculated pattern is shown as a solid line and the measured values are indicated by small circles.

quency range. The standing-wave ratio is less than 2 over a 4.5-to-1 frequency range (700 to 3100 megacycles). As half of the power radiated by the discone is not intercepted by a focus-in-aperture paraboloid, it is to be expected that the gain will be somewhat lower than if a more directive source were used. If the paraboloid is 10 or more wavelengths in diameter, so that the effect of the direct radiation from the discone will be small compared with the total radiation from the paraboloid, the gain relative to an isotropic radiator is[2] approximately

$$G = 0.37\left(\frac{\pi D}{\lambda}\right)^2, \qquad (3)$$

[2] F. E. Terman, "Radio Engineers' Handbook," McGraw-Hill Book Company, New York, New York, 1943, pp. 839–840.

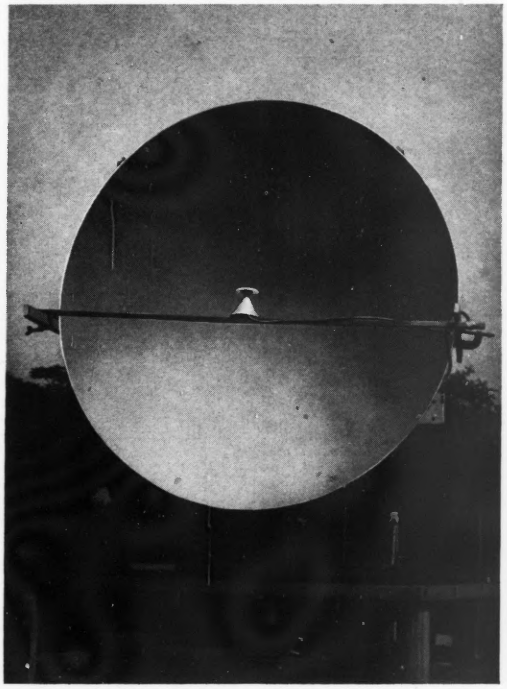


Fig. 11—A single discone radiator and 72-inch parabolic reflector. This arrangement is quite effective where an antenna for a pencil-shaped beam is required to have broad-bandwidth characteristics.

where D = diameter and
λ = wavelength.

This gain is about 4.5 decibels below that of a uniformly irradiated parabolic antenna, and about 2.5 decibels below that of a well-designed narrow-band antenna using a simple dipole and reflector. If the paraboloid is only a few wavelengths in diameter, so that the direct radiation

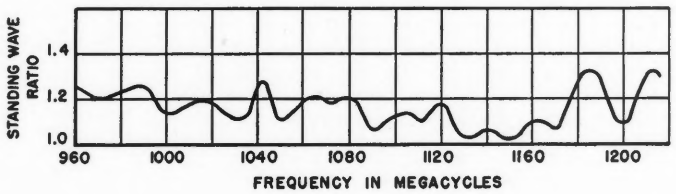


Fig. 10—Standing-wave ratio plotted against frequency. Measured through 50 feet of RG-14/U cable on a 50-ohm slotted line.

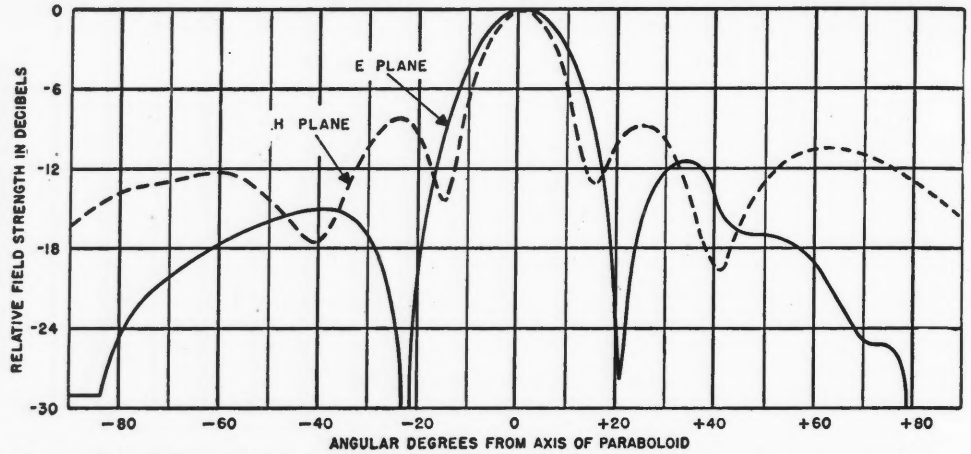


Fig. 12—The solid line indicates the *E*-plane (vertical), and the dashed line indicates the *H*-plane (horizontal) beam pattern of the antenna of Fig. 11. The measurements were made at 600 megacycles.

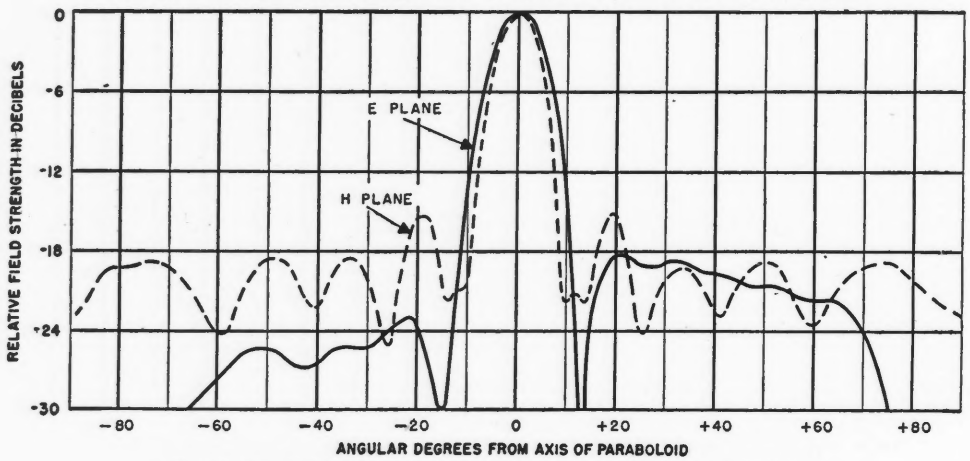


Fig. 13—Parabolic-antenna beam patterns measured at 1290 megacycles.

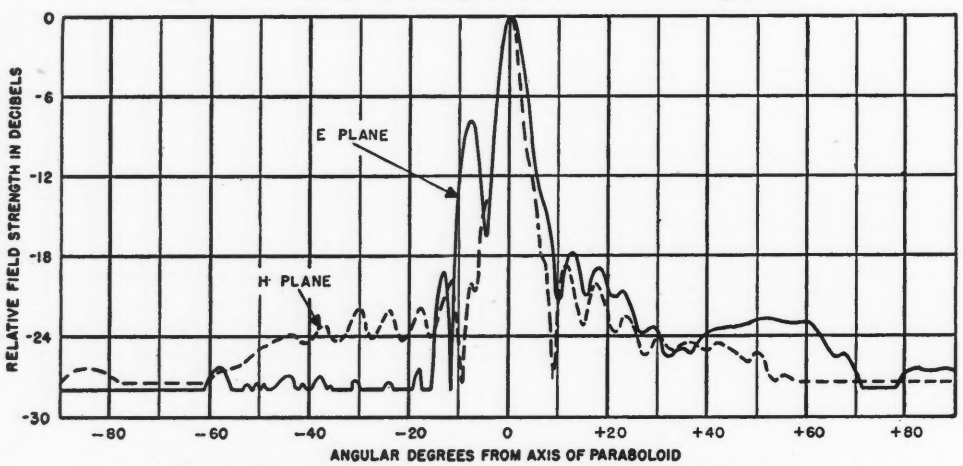


Fig. 14—Parabolic-antenna beam patterns measured at 3200 megacycles.

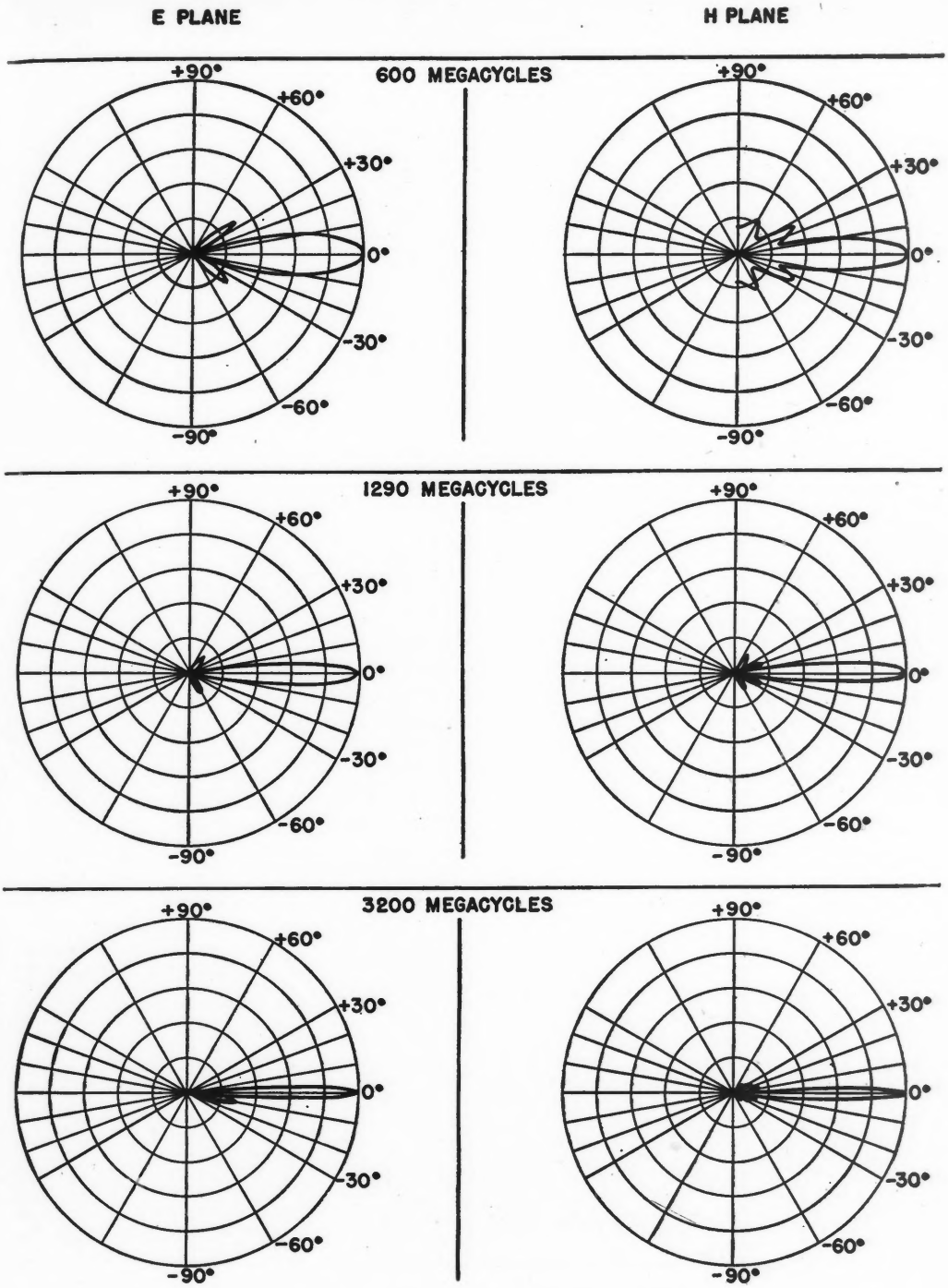


Fig. 15—*E*- and *H*-plane patterns of Figs. 12, 13, and 14 plotted on polar coordinates.

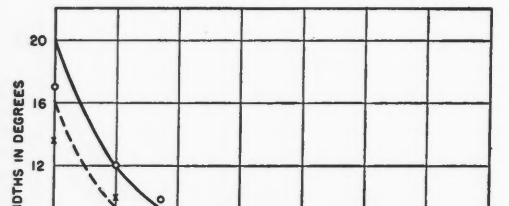


Fig. 16—Variation of the half-power beam widths with frequency for the antenna of Fig. 11. The circles indicate measured results for the *E*-plane pattern; the small crosses, the results for the *H* plane. The solid line indicates the theoretical widths for the *E* plane, calculated from $\theta_B = 1.27\lambda/D$. The dashed line indicates the *H*-plane widths derived from $\theta_B = \lambda/D$. θ_B is the half-power beam width.

becomes an appreciable portion of the total, the gain will depend on the relative phase between the direct radiation and the reflected radiation; the gain may then be larger or smaller than that given by (3). Equation (3) indicates that a 72-inch-diameter paraboloid will have a gain varying from 16.5 decibels at 700 megacycles to 29.5 decibels at 3200 megacycles.

The discone discussed above was used as a radiation source in a 72-inch-diameter paraboloid having a 21.1-inch focal length. The discone was supported by a piece of wood $\frac{3}{4}$ inch wide by $1\frac{1}{4}$ inches thick, which was fastened across the face of the paraboloid (Fig. 11). The discone itself, like the ones used in the distance-measuring-equipment array, consists of a 60-degree cone $4\frac{1}{4}$ inches high and a $3\frac{3}{4}$-inch disc supported above the cone and separated from it by polyethylene insulation. A standard type-N connector is located inside the cone. The RG-8/U cable, which ran from one edge of the paraboloid to the discone, was attached to the wooden support.

Principal electric-plane and magnetic-plane patterns were measured at frequencies of 600, 1000, 1290, 2040, and 3200 megacycles. Some of these patterns are shown in Figs. 12, 13, and 14. In Fig. 15, these same patterns are plotted, more conventionally, in plan coordinates. The measurements were made with the receiving antenna 70 feet from the paraboloid. Fig. 16 shows both a theoretical and measured plot of half-power beam width with frequency. The theoretical values were obtained from patterns calculated on the basis of physical optics.

The standing-wave ratio was measured in the standard manner with 10 feet of RG-8/U cable between the standing-wave detector and the discone. A 6-inch tapered section was used between the standing-wave detector and the cable fitting. Between 850 and 3200 megacycles, the frequency was varied in steps of approximately 20 or 30 megacycles. Four readings were taken between 520 and 680 megacycles; no readings were taken between 680 and 850 megacycles. The standing-wave ratio is plotted against frequency in Fig. 17. Also given in Fig. 17 is the standing-wave ratio of the discone in free space.

The rapid fluctuation in the standing-wave ratio is probably due to the type-N cable connectors and the variations in the dielectric cable itself. These effects are well known and will not be discussed here. The measured standing-wave ratio is slightly lower than that of the antenna itself, due to the

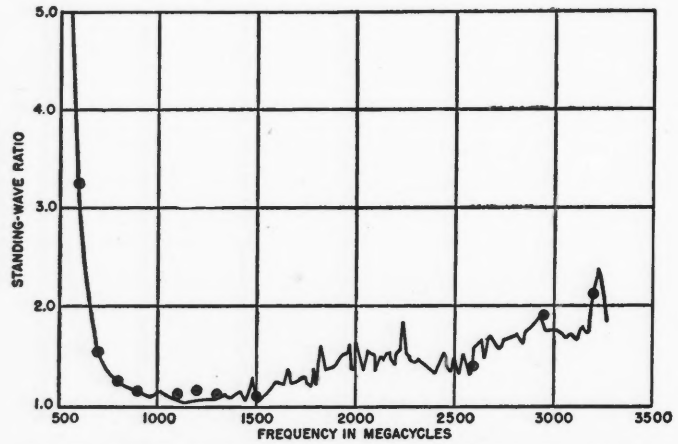


Fig. 17—Standing-wave ratio plotted against frequency for the antenna of Fig. 11. The irregularities are due to the effects of the coaxial cable and associated fittings. Measured results for a discone radiator alone in space are given by the circled points.

cable attenuation. The nominal attenuation of 10 feet of RG-8/U cable is 0.85 decibel at 1000 megacycles, and 1.7 decibels at 3000 megacycles. At 1000 megacycles, a standing-wave ratio of 1.2 will, therefore, be transformed to 1.17, while at 3000 megacycles, a standing-wave ratio of 2.0 will be transformed to 1.74. But in actual operation, some cable must be used, so that it is the standing-wave ratio of the combination that is of interest.

If the focal length of the paraboloid is more than a few wavelengths, the voltage-reflection coefficient introduced into the transmission line by the paraboloid is given[3] by

$$\Gamma = \frac{G\lambda}{4\pi F}, \tag{3}$$

where G = gain of feed in direction of vertex,
λ = wavelength, and
F = focal length.

The gain of the discone being substantially the same as that of a half-wave dipole, this formula gives a reflection coefficient of 0.075 (standing-wave ratio of 1.16) at 1000 megacycles, and a reflection coefficient of 0.025 (standing-wave ratio of 1.05) at 3000 megacycles. This reflection will add to, or subtract from, the reflection of the discone alone, depending on the relative phase between the two reflections. If the reflections add in phase at 2000 megacycles, they will again add in phase at 1750 megacycles, and at 2270 megacycles, assuming that the phase of the discone reflection is constant at these frequencies. This effect is not evident in Fig. 17, but appears to be masked by the other reflections present.

[3] S. Silver, Report No. 442, Radiation Laboratory, Massachusetts Institute of Technology.

Modern Demonstration of MacDonald's Equivalance Theorem

By A. G. CLAVIER

Federal Telecommunication Laboratories, Incorporated, Nutley, New Jersey

A DEMONSTRATION was given in 1912, by MacDonald,[1] of the following proposition, generally known as the "Equivalence Theorem":

The electromagnetic field at any point P outside of a closed surface S due to the movement of electric charges inside S, can be expressed in terms of the electric and magnetic fields produced on S by said electric charges.

MacDonald did not make use of vectorial theory, so that his demonstration was rather lengthy. His method has been followed in the present paper, but considerably shortened by means of suitable notations. An interesting consequence is that it leads directly to the formulas given by Kottler[2] in the case of a harmonic time factor for the contribution of any portion of surface S limited by a closed contour C. It is not necessary to pass through the intermediary of fictitious electric- and magnetic-current densities on surface S.

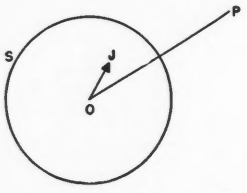


Fig. 1.

Let moving charges be situated inside a closed surface S (Fig. 1), the whole space being homogeneous and isotropic, with inductive capacitances ϵ and μ.

At all points of such a system, the field vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ satisfy the following equations:

$$\operatorname{curl} \boldsymbol{E}+\mu\frac{\partial \boldsymbol{H}}{\partial t}=0, \qquad \operatorname{div} \boldsymbol{H}=0,$$

$$\operatorname{curl} \boldsymbol{H}-\epsilon\frac{\partial \boldsymbol{E}}{\partial t}=J, \qquad \operatorname{div} \boldsymbol{E}=\frac{\rho}{\epsilon}.$$

Factors ρ and J are the charge and current densities at the point considered. They are assumed to be zero for all points situated outside S.

The field vectors at point P, excluding electric fields not varying with time, can be expressed by means of the retarded-vector potential

$$\boldsymbol{A}_P=\frac{\mu}{4\pi}\int_V \frac{1}{r}\,J^*dv,$$

where V is the volume of S. The symbol J^* means that the value of $\boldsymbol{A}$ at time t depends on the value of J at time $t-r/c$ where c is the speed of light in the unbounded medium characterized by the factors ϵ and μ.

Now, r is the distance OP between the point O, where J is applied, and point P at which the field vectors are considered. These vectors satisfy the system

$$\boldsymbol{H}_P=\frac{1}{\mu}\operatorname{curl}_P \boldsymbol{A}_P,$$

$$\epsilon\frac{\partial \boldsymbol{E}_P}{\partial t}=\operatorname{curl}_P \boldsymbol{H}_P=\frac{1}{\mu}\operatorname{curl}_P\cdot\operatorname{curl}_P \boldsymbol{A}_P,$$

where curl_P means that the curl operator implies the coordinates of point P.

The first step of the demonstration consists of expressing the fields at P in terms of the fields at all points O where J is not zero;

$$\boldsymbol{H}_P=\frac{1}{4\pi}\operatorname{curl}_P\int_V\frac{1}{r}\,J^*dv,$$

$$\frac{J^*}{r}=\frac{(\operatorname{curl}_O \boldsymbol{H}_O)^*}{r}-\epsilon\frac{\partial}{\partial t}\frac{\boldsymbol{E}_O^*}{r},$$

$$\frac{J^*}{r}=\operatorname{curl}_O\frac{\boldsymbol{H}_O^*}{r}+\operatorname{curl}_P\frac{\boldsymbol{H}_O^*}{r}-\epsilon\frac{\partial}{\partial t}\frac{\boldsymbol{E}_O^*}{r}.$$

From these equations, the value of $\mu(\partial \boldsymbol{H}_P/\partial t)$ can be derived. Note that

$$\mu\frac{\partial \boldsymbol{H}_O}{\partial t}=-\operatorname{curl} \boldsymbol{E}_O,$$

therefore,

$$\frac{\mu\dfrac{\partial \boldsymbol{H}_O}{\partial t}}{r}=-\frac{(\operatorname{curl} \boldsymbol{E}_O)^*}{r}=-\operatorname{curl}_O\frac{\boldsymbol{E}_O^*}{r}-\operatorname{curl}_P\frac{\boldsymbol{E}_O^*}{r},$$

[1] MacDonald, *Proceedings of the London Mathematical Society*, v. 10, p. 91; 1912.

[2] Kottler, *Annalen der Physik*, v. 71, p. 457; 1923.

so that

$$\mu\frac{\partial H_P}{\partial t}=\frac{\partial}{\partial t}\operatorname{curl}_P\frac{\mu}{4\pi}\int_V\operatorname{curl}_O\frac{H_O^*}{r}dv$$
$$-\operatorname{curl}_P\cdot\operatorname{curl}_P\frac{1}{4\pi}\int_V\operatorname{curl}_O\frac{E_O^*}{r}dv$$
$$-\frac{1}{4\pi}\operatorname{curl}_P\int_V\left(\operatorname{curl}_P\cdot\operatorname{curl}_P\frac{E_O^*}{r}+\epsilon\mu\frac{\partial^2}{\partial t^2}\frac{E_O^*}{r}\right)dv.$$

The third part of the left-hand side of this equation is, however, equal to zero. This is shown by rewriting

$$-\frac{1}{4\pi}\int_V\operatorname{curl}_P\left(\operatorname{grad}_P\operatorname{div}_P-\Delta_P+\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\frac{E_O^*}{r}dv$$
$$=\frac{1}{4\pi}\operatorname{curl}_P\int_V\left(\Delta_P-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\frac{E_O^*}{r}dv.$$

Each component of this vector is easily found to be zero.

Let

$$A_v=\frac{\mu}{4\pi}\int_V\operatorname{curl}_O\frac{H_O^*}{r}dv$$

and

$$A_v'=\frac{-\epsilon}{4\pi}\int_V\operatorname{curl}_O\frac{E_O^*}{r}dv,$$

then,

$$\frac{\partial H_P}{\partial t}=\frac{1}{\mu}\operatorname{curl}_P\frac{\partial A_v}{\partial t}+c^2\operatorname{curl}_P\cdot\operatorname{curl}_P A_v',$$

and disregarding any static distribution of H_P and E_P;

$$\frac{\partial E_P}{\partial t}=-\frac{1}{\epsilon}\operatorname{curl}_P\frac{\partial A_v'}{\partial t}+c^2\operatorname{curl}_P\cdot\operatorname{curl}_P A_v.$$

It is thus seen tht the variable fields E_P and H_P at point P can be expressed in terms of the fields inside S by means of the two vectors A_v and A_v'. It is now a simple matter to transform the volume integrals into surface integrals;

$$\int_V\operatorname{curl}_O\frac{H_O^*}{r}dv=\int_S\left(N^O\times\frac{H_O^*}{r}\right)ds.$$

Let

$$A_S=\frac{\mu}{4\pi}\int_S\left(N^O\times\frac{H_O^*}{r}\right)ds$$

and

$$A_S'=-\frac{\epsilon}{4\pi}\int_S\left(N^O\times\frac{E_O^*}{r}\right)ds,$$

then,

$$\frac{\partial H_P}{\partial t}=\frac{1}{\mu}\operatorname{curl}_P\frac{\partial A_S}{\partial t}+c^2\operatorname{curl}_P\cdot\operatorname{curl}_P A_S',$$
$$\frac{\partial E_P}{\partial t}=-\frac{1}{\epsilon}\operatorname{curl}_P\frac{\partial A_S'}{\partial t}+c^2\operatorname{curl}_P\cdot\operatorname{curl}_P A_S.$$

The computation of the field vectors at point P can obviously be made in two different ways: The first starts with the determination of vectors A_S and A_S'. When this is done, the operations curl_P and $\operatorname{curl}_P\cdot\operatorname{curl}_P$ are performed; these operations involve derivatives with respect to the coordinates of point P only.

A second manner consists of interchanging the processes of integration and differentiation, writing

$$\frac{\partial H_P}{\partial t}=\frac{1}{\mu}\frac{\partial}{\partial t}\frac{\mu}{4\pi}\int_S\operatorname{curl}_P\left(N^O\times\frac{H_O^*}{r}\right)ds$$
$$+c^2\frac{\epsilon}{4\pi}\int_S\operatorname{curl}_P\cdot\operatorname{curl}_P\left(N^O\times\frac{E_O^*}{r}\right)ds,$$

and the corresponding expression for $\partial E_P/\partial t$. This is the same as attributing

$$\frac{1}{4\pi}\left[\frac{\partial}{\partial t}\operatorname{curl}_P\left(N^O\times\frac{H_O^*}{r}\right)\right.$$
$$\left.+\frac{1}{\mu}\operatorname{curl}_P\cdot\operatorname{curl}_P\left(N^O\times\frac{E_O^*}{r}\right)\right]$$

as the contribution of ds to the magnetic field, and

$$\frac{1}{4\pi}\left[-\frac{\partial}{\partial t}\operatorname{curl}_P\left(N^O\times\frac{E_O^*}{r}\right)\right.$$
$$\left.+\frac{1}{\epsilon}\operatorname{curl}_P\cdot\operatorname{curl}_P\left(N^O\times\frac{H_O^*}{r}\right)\right]$$

as the contribution of the same surface element to the electric field.

Now consider the operation $\operatorname{curl}_P\cdot\operatorname{curl}_P=\operatorname{grad}_P\cdot\operatorname{div}_P-\Delta_P$. The first term can be written as follows:

$$\operatorname{grad}_P\cdot\operatorname{div}_P\left(N^O\times\frac{E_O^*}{r}\right)ds$$
$$=\operatorname{grad}_P\frac{E_O^*}{r}\cdot\operatorname{curl}N^O\cdot ds-\operatorname{grad}\cdot N^O\cdot\operatorname{curl}_P\frac{E_O^*}{r}ds.$$

Now, $\operatorname{curl}_P N^O\cdot ds=dl$, the last being the element of the contour integral along the closed contour limiting any portion of surface S.

When there are no line discontinuities of the fields on the closed surface S, the integral $\int_C\operatorname{grad}_P\frac{E_O^*}{r}\cdot dl$ obviously disappears. It must

remain, however, in the correct expression of the contribution of any portion ΔS of surface S, which is thus found to be

$$\frac{\partial \boldsymbol{H}_P}{\partial t}=\frac{\partial}{\partial t}\frac{1}{4\pi}\int_{\Delta S}\operatorname{curl}_P\left(N^0\times\frac{\boldsymbol{H}_O^*}{r}\right)ds$$
$$-\frac{1}{\mu}\frac{1}{4\pi}\int_{\Delta S}\operatorname{grad} N^0\cdot\operatorname{curl}_P\frac{\boldsymbol{E}_O^*}{r}\cdot ds$$
$$-\epsilon\cdot\frac{1}{4\pi}\frac{\partial^2}{\partial t^2}\int_{\Delta S}\left(N^0\times\frac{\boldsymbol{E}_O^*}{r}\right)ds+\frac{1}{4\pi}\int_C\frac{\boldsymbol{E}_O^*}{r}\cdot dl.$$

A similar expression is readily obtained for $\partial \boldsymbol{E}_P/\partial t$. It is only necessary to interchange $\boldsymbol{E}_O^*$ and $\boldsymbol{H}_O^*$, as well as μ and $-\epsilon$.

These formulas have been given by Kottler[2] in the case of a time harmonic function. They were arrived at by means of the intermediary distribution of fictitious electric and "magnetic" current densities on the portion of S considered, and it is then necessary to ascertain that these expressions verify Maxwell's equations at point P. They are, in the present paper, directly derived from those equations and are obviously valid without any further verification.

The "Equivalence Theorem" is found useful in a number of problems. It shows that the experimental determination of the field distribution on any closed surface gives a means of computing the vector fields produced at any external point by sources that are located inside the surface. The same is true when theoretical considerations enable the determination of the fields on a surface S, or portion of said surface; this is generally an approximation only. This last has been found sufficient for the determination of radiation patterns of horns or reflectors used in conjunction with centimeter electromagnetic wavelengths.

Inversely, when the field distribution of secondary sources is assumed to be known on a closed surface S, said sources can be replaced by others that would give the same field distribution on S.

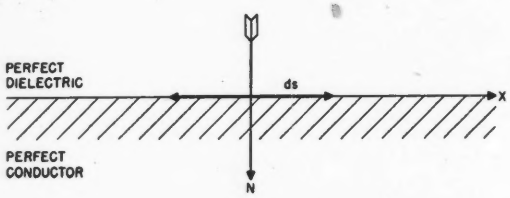


Fig. 2.

Consider, for instance, an unlimited plane supposedly a perfect conductor. Let an electromagnetic wave fall on an element ds of the plane

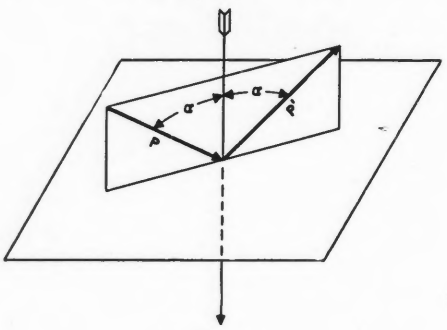


Fig. 3.

in a medium supposed to be a perfect dielectric (Fig. 2). The magnetic inductive capacitance is the same in both media. Let $\boldsymbol{E}$, $\boldsymbol{H}$ be the incoming fields, and $\boldsymbol{E}'$, $\boldsymbol{H}'$ the fields due to the secondary sources induced in the reflecting plane. The boundary conditions give:

$$E_x'=-E_x,$$
$$E_y'=-E_y,$$
$$H_z'=-H_z.$$

There is no dissipation of energy: The flow of the incoming Poynting vector through ds should be equal to the flow of the Poynting vector associated with the reflected wave. It is easy to show that this leads to

$$H_y'=H_y,$$
$$H_x'=H_x.$$

Finally, both incoming and total fields, and consequently both incoming and reflected waves, verify Maxwell's equations. This can be shown to give

$$E_z'=E_z.$$

The association of the above conditions means that the total field $\mathcal{E}$, $\mathcal{H}$ on the reflector is such that

$$\mathcal{E}_z=2E_z,$$
$$\mathcal{E}_T=0,$$
$$\mathcal{H}_z=0,$$
$$\mathcal{H}_T=2H_T.$$

It also means that the incoming and reflected Poynting vectors are in the same plane and make an equal angle with the normal to the plane (Fig. 3).

The secondary sources can then be replaced by fictitious sources that would produce the fields E', H' at all points of the reflecting plane. This obviously leads to the well-known use of "electromagnetic images" to compute the total field in front of the unlimited reflector.

The expressions for the fields at point P given by the "Equivalence Theorem" are, of course, simpler when a time-harmonic function $e^{-i\omega t}$ is assumed. All vectorial operations with index P are then applied to an auxiliary function

$$\varphi = \frac{e^{-i\omega r/c}}{r}.$$

When the radiated fields are considered exclusively, only the terms in $1/r$ need be computed, and $1/r$ can be conventionally treated as a constant in the differentiations involved in the vectorial operations.

$TM_{0,1}$ Mode in Circular Wave Guides with Two Coaxial Dielectrics*

By SIDNEY FRANKEL

Federal Telecommunication Laboratories, Incorporated, New York, New York

FIELD components for a transverse magnetic wave in a wave guide with two coaxial dielectrics are computed. A typical example is given to show the calculation of guide dimensions to reduce phase velocity to a preassigned value.

• • •

A uniform circular wave guide may be conceived as a device that provides an interaction space between electrons and an electromagnetic field to transfer energy from one to the other. If electrons are injected axially in the guide and the field has an axial electric component, the important interaction is between the electrons and this axial component.

For unimpeded motion, the electrons must travel essentially in a vacuum and with a velocity necessarily less than that of light. On the other hand, the phase velocity of the electromagnetic field in an evacuated circular guide is always greater than the velocity of light. For effective interaction, this field must be slowed down, perhaps to the order of one-tenth or less of the velocity of light. This may be accomplished by surrounding the evacuated axial region of the guide with material of high dielectric constant. For simplicity, circular symmetry may be maintained.

An investigation of the propagation of an electromagnetic field with axial electric component in such a "loaded" guide has been made. Appropriate field equations, which take the necessary boundary conditions into account, are set up and solved. A relationship between the maximum value of axial electric component and transmitted power is given.

1. Solution of Field Equations

1.1 Differential Equations

Fig. 1 shows a cross section of a wave guide. The section of the axial region is a circle of radius a, the constants of the medium being μ and ϵ_2, the permeability and dielectric constants, respectively. The boundary of the guide is a perfect conductor of radius b. In the annular region between these two radii is a material of constants μ and ϵ_1. It is assumed that $\epsilon_1 > \epsilon_2$ always.

Cylindrical coordinates r, ϕ, and x will be used with the x axis coinciding with the axis of the guide and being positive in the direction toward a sink of power. For a transverse magnetic mode, take $H_x = 0$, and, in accordance with the conventional $TM_{0,1}$ mode for uniform dielectrics, specify additionally that $E_\phi = H_r = 0$. If a field satisfying these requirements exists in the guide under consideration, then the appropriate field equations in meter-kilogram-second units for either dielectric are of the following type.

$$\left.\begin{aligned} -\partial H_\phi/\partial x &= j\omega\epsilon E_r, \\ (1/r)(\partial/\partial r)(rH_\phi) &= j\omega\epsilon E_x, \\ (1/r)(\partial E_x/\partial\phi) &= 0, \\ (\partial E_r/\partial x) - (\partial E_x/\partial r) &= -j\omega\mu H_\phi, \\ (1/r)(\partial E_r/\partial\phi) &= 0. \end{aligned}\right\} \quad (1)$$

The subscript has been omitted from ϵ to avoid restricting it to either medium.

According to (1), E_r and E_x, therefore also H_ϕ, are independent of ϕ. The useful equations from (1) are

$$\left.\begin{aligned} \partial H_\phi/\partial x &= -j\omega\epsilon E_r, \\ (1/r)(\partial/\partial r)(rH_\phi) &= j\omega\epsilon E_x, \\ (\partial E_r/\partial x) - (\partial E_x/\partial r) &= -j\omega\mu H_\phi. \end{aligned}\right\} \quad (2)$$

Elimination of E_r and E_x among these equations leads to the following equation in H_ϕ:

$$\frac{\partial^2 H_\phi}{\partial x^2} + \frac{\partial}{\partial r}\left[\frac{1}{r}\frac{\partial}{\partial r}(rH_\phi)\right] = -\omega^2\mu\epsilon H_\phi = -\frac{\omega^2}{c^2}H_\phi, \quad (3)$$

where c is the velocity of propagation in a medium of unbounded extent with constants μ and ϵ.

Assume

$$H_\phi = R(r)e^{-\gamma x}, \quad (4)$$

where γ is the propagation constant in the x

* Reprinted from *Journal of Applied Physics*, v. 18, pp. 650–655; July, 1947.

direction and R is a function of r to be determined. Substitution of (4) in (3) yields

$$\frac{d}{dr}\left[\frac{1}{r}\frac{d}{dr}(rR)\right]+\alpha^2R=0, \tag{5}$$

where

$$\alpha^2=\gamma^2+\omega^2/c^2. \tag{6}$$

Solutions of (5) are of the type

$$R=J_1(\alpha r),\ Y_1(\alpha r), \tag{7}$$

where J_1 and Y_1 are Bessel functions of the first and second kinds, respectively, and of the first order. If $Z_1(\alpha r)$ represents any linear combination of these functions, then

$$H_\phi=Z_1(\alpha r)e^{-\gamma z}. \tag{8}$$

For unattenuated transmission, γ is a pure imaginary; let $\gamma=j\beta$, with β real. Then

$$H_\phi=Z_1(\alpha r)e^{-i\beta z},\quad \alpha=[(\omega/c)^2-\beta^2]^{\frac{1}{2}}. \tag{9}$$

From (2), we get immediately

$$\left.\begin{aligned}E_r&=(\beta/\omega\epsilon)Z_1(\alpha r)e^{-i\beta z},\\ E_z&=(\alpha/j\omega\epsilon)Z_0(\alpha r)e^{-i\beta z},\end{aligned}\right\} \tag{10}$$

where Z_0 is a linear combination of Bessel functions of the first and second kind, zeroth order, and we have made use of the relation[1]

$$\frac{d}{dr}[rZ_1(\alpha r)]=\alpha rZ_0(\alpha r). \tag{11}$$

J_0 and J_1 are finite and continuous everywhere in the complex plane, but Y_0 and Y_1 have infinite discontinuities at $r=0$. These latter functions are therefore inadmissable in the axial region of Fig. 1. The solutions therefore run as follows.

In medium 1,

$$\left.\begin{aligned}H_{\phi1}&=[A_1J_1(\alpha_1r)+B_1Y_1(\alpha_1r)]e^{-i\beta_1z},\\ E_{r1}&=\frac{\beta_1}{\omega\epsilon_1}[A_1J_1(\alpha_1r)+B_1Y_1(\alpha_1r)]e^{-i\beta_1z},\\ E_{z1}&=\frac{\alpha_1}{j\omega\epsilon_1}[A_1J_0(\alpha_1r)+B_1Y_0(\alpha_1r)]e^{-i\beta_1z}.\end{aligned}\right\}$$

$$a\leqq r\leqq b. \tag{12a}$$

[1] Jahnke and Emde, "Tables of Functions," 3rd edition, B. G. Teubner, Leipzig, 1938; p. 145.

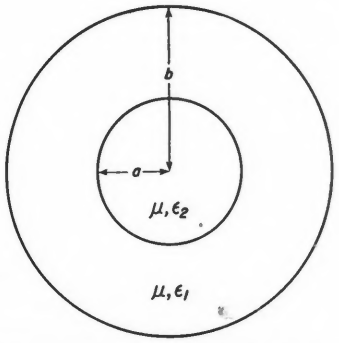


Fig. 1—Cross section of wave guide having two coaxial dielectrics.

In medium 2,

$$\left.\begin{aligned}H_{\phi2}&=A_2J_1(\alpha_2r)e^{-i\beta_2z},\\ E_{r2}&=(\beta_2/\omega\epsilon_2)A_2J_1(\alpha_2r)e^{-i\beta_2z},\\ E_{z2}&=(\alpha_2/j\omega\epsilon_2)A_2J_0(\alpha_2r)e^{-i\beta_2z},\end{aligned}\right\}\ 0\leqq r\leqq a. \tag{12b}$$

A_1, B_1, and A_2 are constants to be determined from the boundary and power conditions, α_1 and α_2 are the propagation constants, and β_1 and β_2 are the phase constants, in media 1 and 2, respectively.

1.2 Boundary Conditions

Matching of tangential components of the electric field at the dielectric interface requires that

$$E_{z1}=E_{z2}\quad\text{at}\quad r=a$$

for all values of x; i.e.,

$$\frac{\alpha_1}{j\omega\epsilon_1}[A_1J_0(\alpha_1a)+B_1Y_0(\alpha_1a)]e^{-i\beta_1z}=\frac{\alpha_2}{j\omega\epsilon_2}A_2J_0(\alpha_2a)e^{-i\beta_2z}. \tag{13}$$

This can be satisfied for all values of x if, and only if, $\beta_1=\beta_2=\beta$. The phase constant (hence the velocity of propagation) is the same in both media, as might have been expected.

$$\alpha_1^2-\omega^2/c_1^2=-\beta^2=\alpha_2^2-\omega^2/c_2^2. \tag{14}$$

Assume the phase velocity V_p to be intermediate between c_1 and c_2

$$c_1<V_p<c_2,\quad V_p=\omega/\beta. \tag{15}$$

This leads to the inequality

$$\alpha_2^2 < 0 < \alpha_1^2, \tag{16}$$

whence α_1 is real while α_2 is a pure imaginary. The field in medium 2 therefore behaves according to Bessel functions of the first kind with pure imaginary argument. Curves for $J_0(jz)$ and $-jJ_1(jz)$, frequently designated as $I_0(z)$ and $I_1(z)$, respectively, have been published.[2] The function $I_0(z)$ is particularly interesting since it shows the radial variation of the axial electric field. This field is fairly uniform for sufficiently small z. For $z<0.6$, the variation is less than 10 percent; for $z<2$, the variation is less than about 2:1.

Next, for the tangential component of electric field to vanish at the conductor surface, it is required that $E_{z1}=0$ at $r=b$, i.e.,

$$A_1 J_0(\alpha_1 b) + B_1 Y_0(\alpha_1 b) = 0. \tag{17}$$

Matching the normal components of dielectric flux at the dielectric interface requires that

$$\epsilon_1 E_{r1} = \epsilon_2 E_{r2} \quad \text{at} \quad r=a; \quad \text{i.e.,}$$

$$A_1 J_1(\alpha_1 a) + B_1 Y_1(\alpha_1 a) = A_2 J_1(\alpha_2 a). \tag{18}$$

Matching the tangential magnetic field at the dielectric interface yields no further information. Collecting and rearranging (13), (17), and (18), we have the set of linear homogeneous equations in A_1, B_1, and A_2:

$$\left.\begin{aligned} \frac{\alpha_1}{\epsilon_1} J_0(\alpha_1 a) A_1 + \frac{\alpha_1}{\epsilon_1} Y_0(\alpha_1 a) B_1 - \frac{\alpha_2}{\epsilon_2} J_0(\alpha_2 a) A_2 &= 0, \\ J_0(\alpha_1 b) A_1 + Y_0(\alpha_1 b) B_1 &= 0, \\ J_1(\alpha_1 a) A_1 + Y_1(\alpha_1 a) B_1 - J_1(\alpha_2 a) A_2 &= 0 \end{aligned}\right\} \tag{19}$$

A necessary and sufficient condition that these equations be consistent is that the determinant of the coefficients vanish:

$$\begin{vmatrix} \frac{\alpha_1}{\epsilon_1} J_0(\alpha_1 a) & \frac{\alpha_1}{\epsilon_1} Y_0(\alpha_1 a) & -\frac{\alpha_2}{\epsilon_2} J_0(\alpha_2 a) \\ J_0(\alpha_1 b) & Y_0(\alpha_1 b) & 0 \\ J_1(\alpha_1 a) & Y_1(\alpha_1 a) & -J_1(\alpha_2 a) \end{vmatrix} = 0. \tag{20}$$

[2] Page 224 of reference 1.

If β is specified in advance, then α_1 and α_2 are known from (14), and (20) specifies a necessary and sufficient relationship between a and b. The manner of using this information will be illustrated by a later example.

With (20) satisfied, any two of the three relations given in (19) may be used to solve for the ratios of the constants A_1, B_1, and A_2. By using the first and the third,

$$A_1 = K_A A_2, \quad B_1 = K_B A_2, \tag{21}$$

where

$$\left.\begin{aligned} K_A &= -\frac{\pi\epsilon_1 a}{2} \begin{vmatrix} \frac{\alpha_2}{\epsilon_2} J_0(\alpha_2 a), & \frac{\alpha_1}{\epsilon_1} Y_0(\alpha_1 a) \\ J_1(\alpha_2 a), & Y_1(\alpha_1 a) \end{vmatrix} \\ K_B &= -\frac{\pi\epsilon_1 a}{2} \begin{vmatrix} \frac{\alpha_1}{\epsilon_1} J_0(\alpha_1 a), & \frac{\alpha_2}{\epsilon_2} J_0(\alpha_2 a) \\ J_1(\alpha_1 a), & J_1(\alpha_2 a) \end{vmatrix} \end{aligned}\right\} \tag{22}$$

and we have made use of the standard identity

$$Y_0(s)J_1(s) - Y_1(s)J_0(s) = 2/\pi s.$$

The field components are then given as follows:

Medium 1: $a \leqq r \leqq b$.

$$\left.\begin{aligned} E_{z1} &= A_2 \frac{\alpha_1}{j\omega\epsilon_1} [K_A J_0(\alpha_1 r) + K_B Y_0(\alpha_1 r)] e^{-j\beta z}, \\ E_{r1} &= A_2 \frac{\beta}{\omega\epsilon_1} [K_A J_1(\alpha_1 r) + K_B Y_1(\alpha_1 r)] e^{-j\beta z}, \\ H_{\phi 1} &= A_2 [K_A J_1(\alpha_1 r) + K_B Y_1(\alpha_1 r)] e^{-j\beta z}, \end{aligned}\right\} \tag{23a}$$

Medium 2: $0 \leqq r \leqq a$.

$$\left.\begin{aligned} E_{z2} &= A_2 \frac{\alpha_2}{j\omega\epsilon_2} J_0(\alpha_2 r) e^{-j\beta z}, \\ E_{r2} &= A_2 \frac{\beta}{\omega\epsilon_2} J_1(\alpha_2 r) e^{-j\beta z}, \\ H_{\phi 2} &= A_2 J_1(\alpha_2 r) e^{-j\beta z}. \end{aligned}\right\} \tag{23b}$$

In (23), the constant A_2 remains to be determined.

1.3 Axial Field as a Function of Transmitted Power

The constant A_2 may be evaluated in terms of the peak transmitted power P_M or twice the average power P. For this purpose, the complex Poynting vector is integrated over a convenient cross section of the guide. This reduces to

$$P_M = \int_0^b \int_0^{2\pi} E_r H_\phi^* \cdot r d\phi dr,$$

where H_ϕ^* is the complex conjugate of H_ϕ. By (22), K_A and K_B both contain the factor α_2 and are therefore pure imaginaries. If E_r and H_ϕ are specified to have zero phase angle at $x=0$, then A_2 is also a pure imaginary, and j appears only in the exponential factor. The integration then runs

$$P_M = 2\pi A_2^2 \frac{\beta}{\omega\epsilon_1} F(a,b;\epsilon_1,\epsilon_2),$$

where

$$F = \frac{\epsilon_1}{\epsilon_2}\int_0^a r J_1^2(\alpha_2 r)dr + \int_a^b r[K_A J_1(\alpha_1 r) + K_B Y_1(\alpha_1 r)]^2 dr \tag{24}$$

or

$$A_2 = ((P_M/2\pi)(\omega\epsilon_1/\beta F))^{\frac{1}{2}}. \tag{25}$$

At the center of the tube, the maximum value of the axial field E_{x2} is

$$E_{x0} = \left| A_2 \frac{\alpha_2}{\omega\epsilon_2} \right| = \left| \frac{\alpha_2}{\omega\epsilon_2} \left(\frac{P_M}{2\pi} \frac{\omega\epsilon_1}{\beta F} \right)^{\frac{1}{2}} \right|. \tag{26}$$

1.4 Further Discussion of the Solution

The determinant (20) may be considered to yield a solution for b when α_1 and α_2 are known. After defining a set of functions

$$Q_1(s) = \frac{J_0(s)}{Y_0(s)}, \quad Q_2(s) = -j\frac{J_1(js)}{J_0(js)}, \quad Q_3(s) = \frac{J_0(s)}{J_1(s)}, \quad Q_4(s) = \frac{Y_1(s)}{J_1(s)}, \tag{27}$$

we can reduce (20) to the form

$$Q_1(\alpha_1 b) = \frac{1 - j\frac{\alpha_1}{\alpha_2}\frac{\epsilon_2}{\epsilon_1} Q_2(\alpha_2 a) Q_3(\alpha_1 a)}{Q_4(\alpha_1 a) - j\frac{\alpha_1}{\alpha_2}\frac{\epsilon_2}{\epsilon_1} Q_2(\alpha_2 a)\frac{Q_3(\alpha_1 a)}{Q_1(\alpha_1 a)}}. \tag{28}$$

All of these Q functions, with the exception of $Q_2(s)$, have been plotted.[3] $Q_2(s)$ is plotted in Fig. 2 of this paper. The use of these curves simplifies the calculation of b.

The chief remaining problem is an evaluation of the function $F(a, b; \epsilon_1, \epsilon_2)$ in (24). The function is integrable in known Bessel functions. By reference to a standard treatise on Bessel

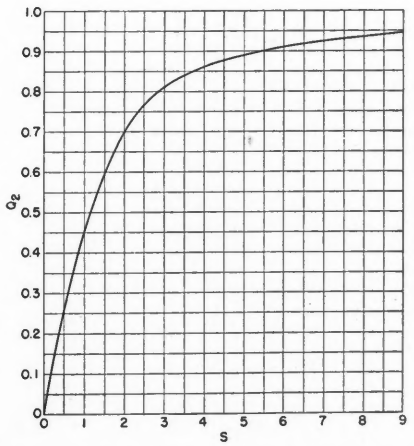


Fig. 2—A useful ratio of certain Bessel functions of imaginary arguments. $Q_2 = -j[J_1(js)/J_0(js)]$.

[3] Pages 200–203 of reference 1.

functions,[4] it can be shown that

$$F=\frac{1}{\alpha_2^2}F_1(\alpha_2 a)+\frac{K_A^2}{\alpha_1^2}[F_1(\alpha_1 b)-F_1(\alpha_1 a)]+\frac{2K_AK_B}{\alpha_1^2}[F_3(\alpha_1 b)-F_3(\alpha_1 a)]+\frac{K_B^2}{\alpha_1^2}[F_2(\alpha_1 b)-F_2(\alpha_1 a)], \quad (29)$$

where

$$\left.\begin{aligned} F_1(s)&=\frac{s^2}{2}\{[J_1(s)]^2-J_0(s)J_2(s)\},\\ F_2(s)-F_2(u)&=\frac{s^2}{2}\{[Y_1(s)]^2-Y_0(s)Y_2(s)\}-\frac{u^2}{2}\{[Y_1(u)]^2-Y_0(u)Y_2(u)\},\\ F_3(s)&=\tfrac{1}{2}\{s^2[Y_1(s)J_1(s)+Y_2(s)J_2(s)]-s[Y_1(s)J_2(s)+Y_2(s)J_1(s)]\}. \end{aligned}\right\} \quad (30)$$

Future work may indicate whether or not these results can be simplified.

2. *Typical Example*

Suppose a $TM_{0,1}$ wave is to be set up at $\omega=3\cdot10^{10}$ per second traveling at one-tenth the velocity of light in vacuum, so that $V_p=3\cdot10^7$ meters per second, whence $\beta=(10)^3$ per meter. Assume the dielectric material is such that

$\mu=4\pi10^{-7}$ henry per meter,
$\epsilon_1=(10^{-7}/18\pi)$ farad per meter
(relative dielectric constant = 200),
$\epsilon_2=(10^{-9}/36\pi)$ farad per meter.

Then $\alpha_1=10^3$, $\alpha_2=j10^3$ by (14). Next, let us try using a radius $a=2$ millimeters $=2\cdot10^{-3}$ meter. Then

$$\alpha_1 a=2,$$
$$\alpha_2 a=j2,$$
$$\alpha_1/\epsilon_1=5.55\cdot10^{11},$$
$$\alpha_2/\epsilon_2=j1.11\cdot10^{14}.$$

From Fig. 2, $Q_2(j2)=0.70$, and from reference 1 and (27),

$$Q_1(2)=0.45,$$
$$Q_3(2)=0.39,$$
$$Q_4(2)=-0.19.$$

Then by (28), $Q_1(\alpha_1 b)=-5.2$ whence

$$\alpha_1 b=0.73,\ 3.78,\ 6.89,\ \ldots.$$

and

$$b=0.7,\ 3.8,\ 6.9,\ \ldots.,\ \text{millimeters}.$$

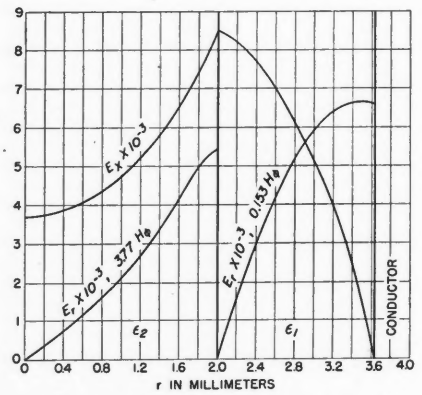


Fig. 3—Distribution of field components in a typical example.

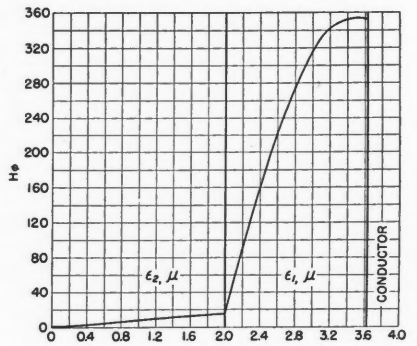


Fig. 4—Distribution of magnetic intensity in a typical example.

[4] G. N. Watson, "A Treatise on the Theory of Bessel Functions," 2nd edition, Macmillan Company, New York; 1944.

We take $b=3.8$ millimeters as the smallest available value of $b>a$. A more careful calculation using tables rather than curves shows that $b=3.6$ millimeters.

Next, compute K_A and K_B using (22).

$$K_A=j153.3,$$
$$K_B=j812.5.$$

The expressions for the field components (23) then become

$$E_{z1}=A_2 2842[J_0(\alpha_1 r)+5.30\,Y_0(\alpha_1 r)]e^{-j\beta z},$$
$$E_{r1}=jA_2 2892[J_1(\alpha_1 r)+5.30\,Y_1(\alpha_1 r)]e^{-j\beta z},$$
$$H_{\phi 1}=jA_2 153.3[J_1(\alpha_1 r)+5.30\,Y_1(\alpha_1 r)]e^{-j\beta z},$$
$$E_{z2}=A_2 3703 J_0(\alpha_2 r)e^{-j\beta z},$$
$$E_{r2}=A_2 3774 J_1(\alpha_2 r)e^{-j\beta z},$$
$$H_{\phi 2}=A_2 J_1(\alpha_2 r)e^{-j\beta z}.$$

These quantities are plotted as functions of r in Figs. 3 and 4. Here it is plain that the electric vectors have the same orders of magnitude in both dielectrics, while the magnetic vector is practically negligible in the central dielectric. This fact, combined with the small cross-sectional area of the inner dielectric, results in most of the transmitted power being propagated in the outer dielectric.

The factor F in (24) may be evaluated by numerical integration, whence it turns out that $F=-0.00037-0.3475=-0.3479$, the smaller term being the contribution of the inner dielectric. Since by (26) the transmitted power is related linearly to F, it can be seen that only about 0.1 percent of the power is propagated in the central region.

By substitution in (26) we now have immediately $E_{x0}=5.77(P_M)^{\frac{1}{2}}$, volts per centimeter, at the axis of the guide.

Discussion

Remarks on Slow Waves in Cylindrical Guides*

SEVERAL articles, one in 1944 [1] and two more recently, [2] have appeared on the question of concentric dielectrics in circular wave guides, each treating it from a different point of view. The problem considered by the more recent articles is the utilization of the $TM_{0,1}$ mode in a circular wave guide with two coaxial dielectrics to obtain wave propagation with a phase velocity below that of light velocity. A main objective of each of these articles is to obtain a method whereby one is able to determine the necessary values of the dielectric constants and radii involved, when the phase velocity is given some pre-assigned value. The earlier article considers the problem more generally. All of the authors approach the problem by solving for the fields and then applying the appropriate boundary conditions.

The purpose of this letter is twofold; (1) to indicate that an alternative and simpler method may be used to derive the necessary relations, and (2) to acquaint the reader with the existence of certain curves whose use greatly eases the calculations.

The properties of an ideal wave guide are determined once the cut-off frequency is determined. The cut-off frequency for a given mode corresponds to resonance for waves propagating transversely in the cross section. Therefore, to find the cut-off frequency of the guide, it is necessary only to determine the resonant frequency of the two-dimensional problem defined by the guide cross section. The condition for resonance that we apply is that, at any radius, the admittances looking in opposite radial directions should be equal and opposite.

Since the above-mentioned articles do not agree as to notation, it is in place to specify the notation used here. The inner and outer regions

* By A. A. Oliner, Microwave Research Institute, Polytechnic Institute of Brooklyn, Brooklyn, New York, reprinted from *Journal of Applied Physics*, v. 19, pp. 109–110; January, 1948.

[1] L. Pincherle, "Electromagnetic Waves in Metal Tubes Filled Longitudinally with Two Dielectrics," *Physical Review*, v. 66, p. 118; 1944.

[2] Sidney Frankel, "$TM_{0,1}$ Mode in Circular Wave Guides with Two Coaxial Dielectrics," *Journal of Applied Physics*, v. 18, p. 650; 1947; G. G. Bruck and E. R. Wicher, "Slow Transverse Magnetic Waves in Cylindrical Guides," *Journal of Applied Physics*, v. 18, p. 766; 1947.

shall be regions 1 and 2, respectively (see Fig. 1). Dielectric constant ϵ_2 is greater than ϵ_1. k_c shall represent the cut-off wave number, and β, the propagation constant in the axial direction. We then have:

$$k_{c1}=(k^2\epsilon_1-\beta^2)^{\frac{1}{2}}, \quad k_{c2}=(k^2\epsilon_2-\beta^2)^{\frac{1}{2}}, \qquad (1)$$

where $k=\omega/c$ and $\beta=\omega/v_p$, v_p being the phase velocity.

A radial guide is the appropriate one to use in the cross section since we have propagation down a circularly cylindrical wave guide. The expression for the admittance $Y(r_1)$ at radius r_1 in terms of the admittance $Y(r_2)$ at r_2, for the lowest E-type mode in a radial guide can be written in analogy with the uniform line case as:[3]

$$Y(r_1)=Y_0(r_1)\frac{j+\zeta(x,\ y)ct(x,\ y)\,Y(r_2)/Y_0(r_2)}{Ct(x,\ y)+j\zeta(x,\ y)\,Y(r_2)/Y_0(r_2)}, \qquad (2)$$

where

$$\zeta(x,\ y)=\frac{J_0(x)N_0(y)-N_0(x)J_0(y)}{J_1(x)N_1(y)-N_1(x)J_1(y)},$$

$$ct(x,\ y)=\frac{J_1(x)N_0(y)-N_1(x)J_0(y)}{J_0(x)N_0(y)-N_0(x)J_0(y)}, \qquad (3)$$

$$Ct(x,\ y)=\frac{J_1(y)N_0(x)-N_1(y)J_0(x)}{J_1(x)N_1(y)-N_1(x)J_1(y)},$$

$x=k_cr_1$, $y=k_cr_2$, and Y_0 is the characteristic admittance of the guide which is a function of the radius. It may be seen that if a short circuit exists at r_2, then $Y(r_2)=\infty$, and $Y(r_1)$ reduces to:

$$\bar{Y}(r_1)=-jY_0(r_1)ct(x,\ y). \qquad (4)$$

Similarly, we find that for any finite $Y(r_2=0)$:

$$\bar{Y}(r_1)=jY_0(r_1)\frac{J_1(x)}{J_0(x)}. \qquad (5)$$

The condition for resonance is satisfied if, at radius $r=r_1$, we set $\bar{Y}+\bar{Y}=0$. We then obtain:

$$Y_{01}(r_1)\frac{J_1(k_{c1}r_1)}{J_0(k_{c1}r_1)}=Y_{02}(r_1)ct(x,\ y), \qquad (6)$$

where $x=k_{c2}r_1$ and $y=k_{c2}r_2$. For E-type modes the ratio of the characteristic admittances is given by

$$\frac{Y_{01}(r_1)}{Y_{02}(r_1)}=\frac{\epsilon_1k_{c2}}{\epsilon_2k_{c1}}. \qquad (7)$$

We may note that k_c is the propagation constant in the transverse region, although β is the propagation constant in the axial direction. Substituting for the characteristic admit-

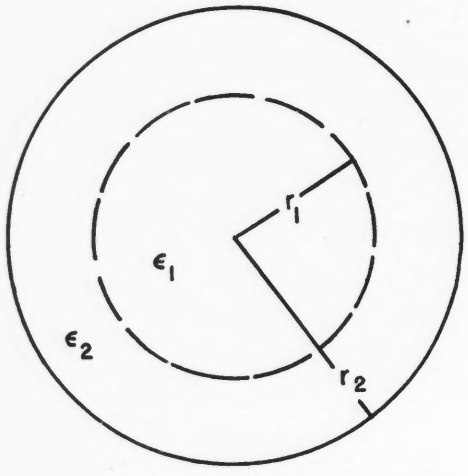


Fig. 1.

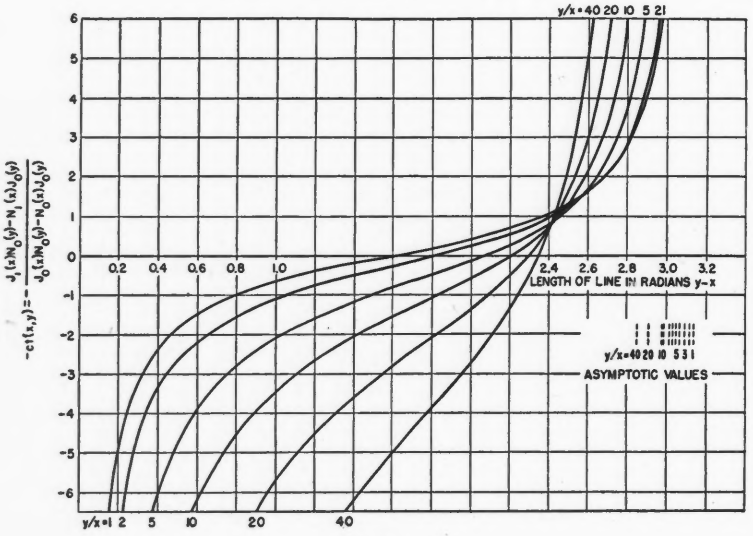


Fig. 2.

[3] Montgomery, Dicke, and Purcell, "Principles of Microwave Circuits" (v. 8, Radiation Laboratory Series), McGraw-Hill Book Company, New York, New York, to be published; a brief derivation may be found in Chapter 8.

tances, we obtain finally:

$$\frac{\epsilon_1 k_{c2}}{\epsilon_2 k_{c1}} \frac{J_1(k_{c1} r_1)}{J_0(k_{c1} r_1)} = ct(x, y), \tag{8}$$

where $x = k_{c2} r_1$ and $y = k_{c2} r_2$.

It may be seen that (8) is identical, except for notation, with (8) of the article by Bruck and Wicher. These authors state that this expression is the relation between the field velocity, tube radii, and dielectric constant, from which they obtain their results. They further state that this relation "can be solved by a combination of patience with a variety of numerical procedures." However, there exist curves of the combination of Bessel functions represented by $ct(x, y)$. (See Fig. 2.) The values of $ct(x, y)$ are plotted against $y-x$, with y/x as a parameter. The results desired may be obtained quite rapidly by the use of these *ct* curves. The method involved here is essentially the same as that used by these authors; the curves simply serve to facilitate the calculations.

As an example, one of the points considered by the above authors was checked using the *ct* curves. The phase velocity chosen was $v_p = 0.2c$, and the smaller and larger radii were 0.05λ and 0.10λ, respectively. The required ratio of dielectric constants must be determined. The value $\epsilon_2/\epsilon_1 = 57.3$ was obtained by means of the *ct* curves. Bruck and Wicher state that $\epsilon_2/\epsilon_1 = 57.25$.

In the article by Frankel, the final relations are obtained in somewhat different form. However, after appropriate rearranging of terms, (28) of Frankel's article may be seen to be identical, again except for notation, with (8) above. In his "typical example," Frankel specifies the values of all the necessary parameters except the larger radius, which must be solved for. His result, obtained graphically, is 3.8 millimeters. The result obtained by means of the *ct* curves is 3.77 millimeters. Frankel states further that "a more careful calculation using tables rather than curves" shows that the result should be 3.6 millimeters. The writer believes that this statement is in error as the use of tables with (8) above yields a value of 3.773 millimeters.

High-Ratio Multivibrator Frequency Divider

By M. SILVER and A. SHADOWITZ

Federal Telecommunication Laboratories, Incorporated, Nutley, New Jersey

FREQUENCY-DIVIDER chains of considerable length for obtaining two widely separated frequencies in synchronism with each other have been increasingly used in recent years. In television practice, for example, it is necessary to generate a 60-cycle-per-second signal in exact lock with another of 31,500 cycles.[1] In frequency-modulation broadcast transmitters of the direct type, it is usual to divide the master-oscillator frequency by a factor of several hundred in order to reduce the index of modulation before comparison with the crystal oscillator. It has been customary to use a comparatively large number of dividers, either of the multivibrator or blocking-oscillator type, to accomplish the above, since each divider must be restricted to a rather small order of division, about 5, for stability under varying operating conditions. A system[2] has also been designed for accomplishing high-order division with only a few stages, by using saw-tooth dividers. The present article describes a somewhat similar circuit for a high-ratio multivibrator divider.

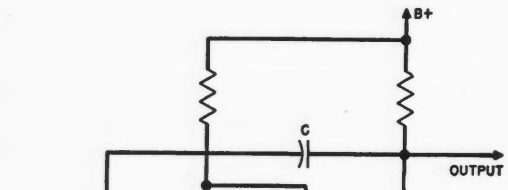


Fig. 1—Typical multivibrator-divider circuit.

[1] A. Rufus Applegarth, "Synchronizing Generators for Electronic Television," *Proceedings of the I.R.E.*, v. 34, pp. 128W–139W; March, 1946.

[2] Geoffrey Builder, "A Stabilized Frequency Divider," *Proceedings of the I.R.E.*, v. 29, pp. 177–181; April, 1941.

1. Circuit Operation

1.1 Simple Lock-In Multivibrator

Fig. 1 is the circuit diagram of an ordinary multivibrator divider. A synchronizing voltage (either pulse or sine wave) applied at the input will lock the multivibrator output frequency at an exact submultiple of the synchronizing frequency if the time constants and the amplitude of the input voltage are properly chosen. It is difficult to stabilize the order of division with this circuit; instead of dividing, say, by 20/1 as desired, a slight change in any one of a number of factors may cause division by 19/1 or 21/1. The value of the plate-supply voltage, the synchronizing-voltage amplitude or frequency, and aging of the tubes all combine to make high-order division critical.

1.2 Insertion of Clippers

Suppose that we remove capacitor C (Fig. 1) between the first grid and second plate, and replace it with a tuned circuit followed by a two-stage clipper, as shown in Fig. 2. The output of the second clipper is fed back to the first grid. Resistor R has a high value in order that the Q of the tuned circuit may be high, and the bias on the first clipper stage is made high enough to prevent the flow of grid current. The capacitors connected with the clippers are all comparatively large.

It can be seen that this circuit will oscillate only in the immediate vicinity of the resonant frequency of the tuned circuit, with or without synchronizing voltage at the input. A sine wave will be obtained at output A, while one side of the sine wave will be clipped at the grid of the second clipper stage. At output B (and in the multivibrator circuit proper), a square wave will be obtained.

If we introduce a synchronizing voltage of a frequency close to an integral multiple of the tuned-circuit frequency, the multivibrator will

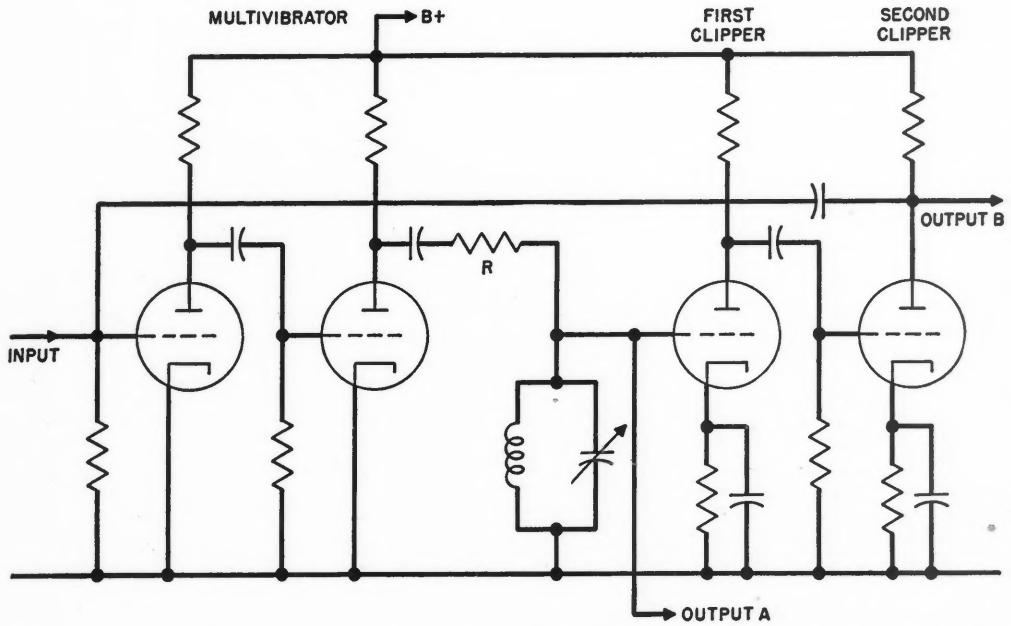


Fig. 2—Modified circuit of Fig. 1, in which capacitor C is replaced by a tuned circuit and two clipper stages. Larger and more stable division ratios between input and output frequency are possible with this circuit.

adjust itself to an exact subharmonic of the input.

The action here is that of a lock-in of an unstable oscillator to a stable oscillator; a necessary condition for lock-in is that the two oscillators should have a harmonic frequency in common, and this is the reason for the double-clipper stage, which generates harmonics by squaring the sine wave. The frequency of the synchronizing voltage (the stable oscillator) must be close to one of the multivibrator harmonics for lock-in to occur; the higher the order of division desired, the more important it is that the waves be square. Since it is difficult to obtain good square waves at high repetition rates, it is also more difficult to obtain good division at the higher frequencies. The actual lock-in forces are produced by third-order nonlinearities in the tube operating characteristic.

1.3 Lock-In Range

For any given order of division, there is a range of synchronizing frequencies at which proper division is obtained. An example is given in Fig. 3. This synchronizing-frequency range is directly proportional to the multivibrator operating range, which is determined by the Q of the tuned

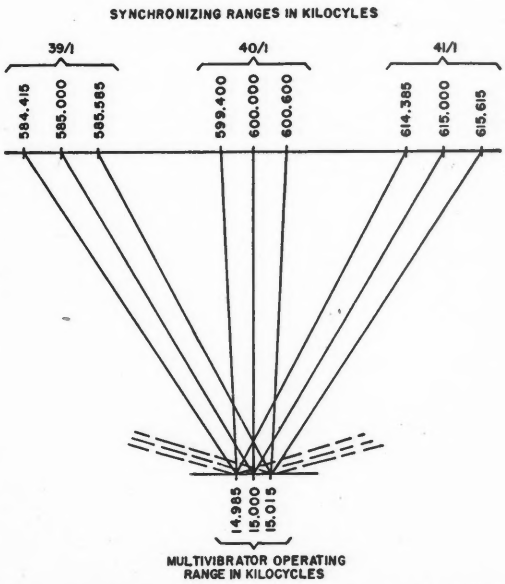


Fig. 3—Illustration of three of the synchronizing ranges of the multivibrator-divider circuit of Fig. 2.

circuit. If the multivibrator is driven at a different frequency from the exact resonant frequency of the tuned circuit, a phase shift is introduced

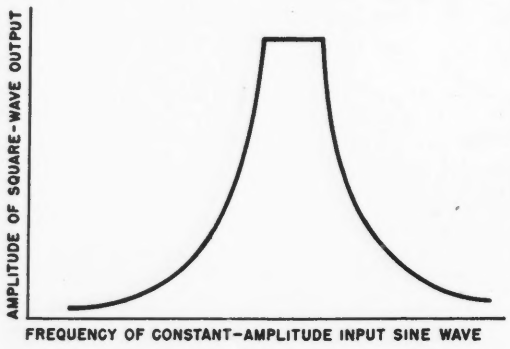


Fig. 4—Transmission characteristic of tuned-circuit-and-clipper coupling circuit. It exhibits the typical frequency-versus-amplitude curve of a tuned circuit, with a plateau at the top due to the clippers.

in the tuned circuit; the phase shift for a given deviation in frequency increases with increasing Q, and proper division no longer occurs when the phase shift exceeds a certain critical value. Thus for higher orders of division, where adjacent synchronizing ranges become closer to each other, it is necessary to restrict the multivibrator operating range to smaller regions by using higher Q's.

The transmission characteristic of the circuit is shown in Fig. 4. This curve may be obtained by removing the capacitor between the input grid and the second clipper plate, applying a variable-frequency constant-amplitude sine wave to the input, and examining the output at the second clipper plate. The output voltage will reach the maximum value at very low input voltages (of the order of millivolts), and will then be a square wave. The plateau width (Fig. 4) can be varied by changing the bias of the first clipper, and bears no relation to the multivibrator lock-in range.

2. Final Circuit

Fig. 5 is the detailed schematic of a circuit that gives stable division ratios as high as 300/1 (4.5 megacycles to 15 kilocycles). Adjustment for proper division can be made most easily with an oscilloscope at any of the plates; when the multivibrator locks with the synchronizing frequency, it is possible to count the number of synchronizing-frequency pulses per multivibrator cycle.

3. Conclusion

The circuit described has the one outstanding advantage of simplicity; it requires but one tuned circuit and two tubes, and is quite insensitive to changes in operating conditions. In fact, since the values of resistance and capacitance have so little effect on the frequency, it would seem incorrect to call the circuit a multivibrator, except that it is derived from the ordinary multivibrator, and because it produces easily synchronized square waves.

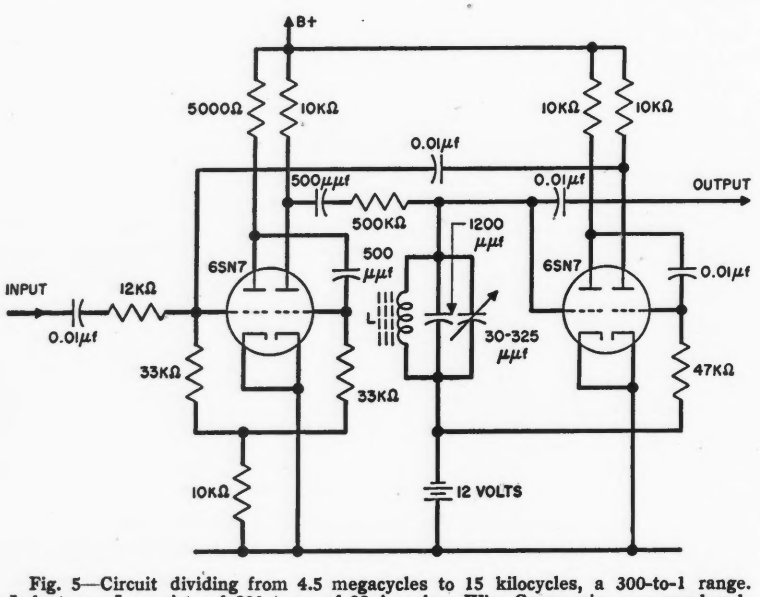


Fig. 5—Circuit dividing from 4.5 megacycles to 15 kilocycles, a 300-to-1 range. Inductance L consists of 800 turns of 28-American-Wire-Gauge wire on a powdered-iron toroid.

Thermo-Electric and Conductive Properties of Blue Titanium Dioxide*

By H. K. HENISCH

Department of Physics, University of Reading, England

1. Introduction

THE thermo-electric properties of blue semi-conducting titanium dioxide are of special interest, both in connection with the possible practical applications of this material and because of the theoretical importance of these effects. Titanium dioxide is also in many ways an ideal semi-conductor because its conductivity is purely electronic up to quite high temperatures[1] and the material is entirely stable under all normal conditions. Moreover, as far as is known, its structure with regard to electronic-energy distribution is substantially in accordance with the simple concepts on which semi-conductor theories are usually based. Experiments on this material should, therefore, offer welcome opportunities for checking the validity of these theories.

Most theoretical work on semi-conductors necessarily assumes that the properties concerned are those of a single crystal of the material. During practical experiments, this condition is rarely satisfied. Some experiments have indeed been carried out on single crystals and have yielded results of great value,[2,3] but most semi-conductors otherwise available for investigation are in the form of micro-crystalline conglomerates. This means that properties like the conductivity will depend not only on the characteristics of individual grains, but also on the shape, orientation, and packing of grains within the conglomerate as a whole. As a rule, it is not possible to control these two contributing factors independently during the preparation of specimens, nor is it possible to distinguish between them by conductivity measurements alone. To test the predictions of theory, it is, therefore, necessary to measure some property that depends only on the composition (impurity content) and not on the grain structure. The thermo-electric and Hall effects are two such properties and of these the former is likely to be the simpler to measure. Thermo-electric measurements are thus of value as means not only of testing the theory of semi-conductors, but also (in combination with conductivity measurements on the same sample) of enabling the relative importance of composition and grain structure to be assessed.

A considerable amount of experimental and theoretical work has been carried out on this subject in the past, probably most of it with special reference to cuprous oxide,[4–13] but some also in connection with selenium,[14] lead sulphide,[15] stannous sulphide,[16] germanium,[17,18] cadmium oxide,[19] silicon,[20] zinc oxide,[21] and many other substances and mixtures.[22] Expressions for the thermo-electric power and the total thermo-electromotive force have been developed in terms of the temperatures T_1 and T_2 of two metal–semi-conductor junctions and various parameters characteristic of the materials concerned.[10,17,23] In practice, many of these equations cannot readily be subjected to experimental test, because they are complicated and contain terms that cannot be independently estimated. It will be shown that by suitable choice of parameters it is possible to deduce approximate equations that are more suitable for direct experimental check, and the application of which leads to results of practical interest.

As is well known, the temperature dependence of the conductivity is given by an equation

$$\sigma = \sigma_0 \exp\left[-eV_a/2kT\right],$$

where σ_0 usually varies much more slowly with temperature than the exponential term. According to the simplest theory, eV_a is the energy required to free an electron from an impurity centre and raise it into the conduction band. This activation energy also enters into the thermo-electric equations. It is shown in the present paper that the experimental values for this energy derived from conductivity and thermo-electric measurements on sintered

* This article arises from work on semi-conductors that is being undertaken by the University of Reading, England, in behalf of Standard Telecommunication Laboratories Limited.

[1] All numbered references will be found on p. 177.

titanium dioxide specimens agree only as regards order of magnitude. This would be expected for specimens that are not entirely homogeneous, in the sense that their impurity content and activation energy are subject to local variations. It is suggested that while the impurity content can vary continuously, only a few discrete values of the activation energy are possible and the different grains each have one or other of these values. The measured activation energy thus represents some form of average. The effect of non-uniform impurity content on the thermo-electric properties of the specimens can be calculated for an ideal case in which the impurity concentration varies continuously throughout the solid. The conclusions are supported by the experimental results and are thought to be qualitatively correct even when the assumption of continuous variation is no longer satisfied.

For a single homogeneous crystal, σ_0 should be proportional to the square root of the concentration of impurity centres.[24] For a micro-crystalline aggregate, this relation is generally not fulfilled, but it is found experimentally that some such correlation between σ_0 and the average concentration of impurity centres (as inferred from thermo-electric measurements) still exists. The correlation is not exact, indicating that the intergranular boundaries and the grains themselves make comparable contributions to the resistivity.

The thermo-electric properties of semi-conductors have a practical as well as a theoretical interest owing to the magnitude of the thermo-powers involved. Thermo-couples consisting of a metal and a semi-conductor can have thermo-electric powers of the order of 1 millivolt per degree centigrade, which is between 10 and 80 times as high as those observed on purely metallic couples. Semi-conductors may thus be very suitable for the construction of sensitive thermo-piles, and indeed have already been used for this purpose.

2. *Thermo-Electric Effects in Homogeneous Semi-Conductors*

To simplify the problem as much as possible, the following treatment concerns only semi-conductors that are either completely of the "excess" or completely of the "deficit" type, and that are not subject to "intrinsic" semi-conduction within the temperature range under discussion.

Fig. 1 shows the approximate energy relations at the contact between a metal and a typical

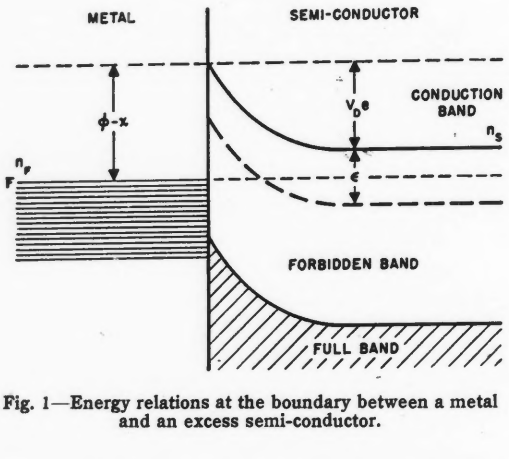


Fig. 1—Energy relations at the boundary between a metal and an excess semi-conductor.

excess semi-conductor. F is the Fermi-level of the metal, ϵ the activation energy of the semi-conductor, ϕ and χ are the thermionic work-functions of the metal and the semi-conductor, respectively, and V_D is called the diffusion potential. The potential barrier is of the type suggested by Schottky, but its precise shape or thickness is of no importance for present considerations. When contact is made, the relative heights of the Fermi-level and the conduction band adjust themselves so that the electrons on either side of the barrier are in dynamic equilibrium, i.e., so that equal numbers of electrons cross the barrier in opposite directions during any given time. This can only be if the electron concentration at any particular level, e.g., the peak of the barrier, is uniform. Since the electron concentrations decrease very nearly exponentially with increasing energy, the necessary equilibrium condition is

$$n_F \exp\left[-(\phi-\chi)/kT\right]=n_S \exp\left[-eV_D/kT\right], \quad (1)$$

where n_F is the electron concentration at the Fermi-level of the metal and n_S that at the bottom of the conduction band in the semi-conductor. If Π is the contact potential at the junction, then

$$e\Pi=\phi-\chi-eV_D \quad (2)$$

and hence, after substitution into (1) and re-arrangement,

$$\Pi = (Tk/e) \log n_F/n_S. \quad (3)$$

This should not be confused with a similar equation sometimes encountered in the literature (and now known to be erroneous), in which n means the *total* electron concentration, independent of energy level.

As a first approximation, we can assume that n_S is given by some expression

$$n_S = A \exp [-\epsilon/2kT], \quad (4)$$

where the "concentration constant" A is nearly independent of temperature compared with the exponential term and is proportional to the square root of the concentration of impurity centres (compare (13)). Thus, we have

$$\Pi = (Tk/e)(\log n_F/A + \epsilon/2kT). \quad (5)$$

n_F is not strictly independent of temperature (it is expected to decrease slightly as T increases), but it is convenient to neglect this for the present, i.e., to assume that the Thomson coefficient of the metal is negligible compared with that of the semi-conductor. Integration of (5) gives

$$E = \int_{T_2}^{T_1} (\Pi/T)dT = (T_1 - T_2)(k/e) \log n_F/A + (\epsilon/2e) \log T_1/T_2, \quad (6A)$$

where E is in absolute units, or

$$E = (T_1 - T_2)(300k/e) \log n_F/A + (V_a/2) \log T_1/T_2, \text{ Volts}, \quad (6B)$$

where V_a is the activation potential in volts. The polarity of the electromotive force is such that the metal electrode at the hot junction becomes positive with respect to that at the cold junction. Equation (6B) already satisfies a number of important conditions. It is symmetrical in T, so that inter-change of T_1 and T_2 merely reverses the sign of the thermo-electromotive force. Similarly, $E=0$ when $T_1=T_2$. Inspection also shows that the law of intermediate metals and the corresponding law of intermediate semi-conductors are obeyed.

The thermo-dielectric power is, of course, given by

$$(dE/dT)_T = (300k/e) \log n_F/A + V_a/2T, \quad (7)$$

according to which dE/dT cannot be zero at any temperature, unless $n_F < A$. Cases in which this condition is definitely satisfied are not known.

Re-arrangement of (6B) leads to

$$\frac{E}{T_1 - T_2} = \frac{300k}{e} \log \frac{n_F}{A} + \frac{V_a}{2}\left(\frac{\log T_1/T_2}{T_1 - T_2}\right). \quad (8)$$

Thus, if the present considerations are correct, a straight line should be obtained by plotting the average thermo-power $E/(T_1 - T_2)$ against the function $(\log T_1/T_2)/(T_1 - T_2)$. The activation energy should be directly deducible from the slope of this line, and its intercept on the ordinate should be a function of the concentration of impurity centres. The average thermo-power is expected to be high for semi-conductors of low impurity content, and vice versa.

The same argument is also applicable to a pure deficit semi-conductor, with the important difference that the resulting thermo-electromotive forces are of the opposite polarity, i.e., the metal electrode at the hot junction becomes negatively charged. It was previously thought possible to determine the character of a given semi-conductor on the basis of this criterion alone, but it will be shown that this is not always reliable, even when the sign of $\log n_F/A$ in (7) is known.

The present treatment is based on the assumption that n_F/A can be regarded as constant. A more accurate treatment would have to take account of the temperature dependence of A and of the finite Thomson coefficient of the metal. To do this with uncompromising precision would introduce considerable complications. It is, however, possible to introduce a second stage of approximation without great difficulty. To do this, it is convenient to put

$$n_F/A = (n'_F/A')T^q, \quad (9)$$

where q is some constant, and to remember that

$$\sigma_s - \sigma_m = \Pi/T - d\Pi/dT, \quad (10)$$

where σ_s and σ_m are the Thomson coefficients of the semi-conductor and the metal respectively. Substitution into the previous equations then leads to

$$\sigma_s - \sigma_m = \epsilon/2eT - qk/e. \quad (11)$$

Hence, if σ_s and σ_m were known independently,

the value of the constant q could be estimated by this indirect method.

The Thomson coefficient of the semi-conductor can be simply calculated as follows. When two parts of a semi-conductor are at different temperatures, the electrons in the conduction band will reach equilibrium under the influence of two opposing tendencies: the normal process of thermal diffusion and the electric field resulting from any charge displacement. Electrons will drift from the hot to the cold regions, because their concentration as well as their kinetic energy increases very greatly with increasing temperature. When equilibrium is reached and no current flows, this tendency to drift must just be counter-balanced by the electric field. The condition for this is

$$ebnF_x = eD(dn/dx), \tag{12}$$

where b is the electronic mobility, F_x the local electric field, D the diffusion coefficient ($=kTb/e$), and n the *total* electron concentration in the conduction band, given by the well-known expression

$$n = CN^{\frac{1}{2}}T^{\frac{3}{4}} \exp\left[-\epsilon/2kT\right], \tag{13}$$

where C is a constant and N the concentration of impurity centres. We then have

$$\left.\begin{aligned} F_x &= (kT/en)(dn/dx) \\ &= (kT/en)(dn/dT)(dT/dx) \end{aligned}\right\} \tag{14}$$

or:

$$(kT/e)(\epsilon/2kT^2+3/4T)(dT/dx) = dE/dx = (dE/dT)(dT/dx). \tag{15}$$

But dE/dT is the Thomson coefficient* σ_s, so that

$$\sigma_s = \epsilon/2eT+3k/4e. \tag{16}$$

Equation (13) is expected to hold with accuracy only as long as the concentration of free electrons is small compared with that of the impurity centres (see page 158 of reference 24), and (16) is, therefore, subject to the same limitation. It is necessary to state that this assumption may not be entirely justified in the case of titanium dioxide.

With regard to σ_m, the first approximation was zero, and as a second approximation $\sigma_m = 3k/2e$ is probably justified, although this is strictly correct only for a non-degenerate electron gas. We thus have

$$\sigma_m = \sigma_s - \epsilon/2eT + qk/e = (k/e)(3/4+q) = 3k/2e, \tag{17}$$

so that

$$q = 3/4. \tag{18}$$

This can now be substituted into (9) and hence into (5), resulting in

$$\Pi/T = (k/e)\log n'_F/A' + (3k/4e)\log T + \epsilon/2eT, \tag{19}$$

which can be integrated with respect to temperature to give in absolute units:

$$E = (T_1-T_2)(k/e)\log n'_F/A' + (\epsilon/2e)\log T_1/T_2 + (3k/4e)\left[T(\log T-1)\right]^{T_1}_{T_2}. \tag{20}$$

After re-arrangement and conversion into volts, this becomes

$$\frac{E}{T_1-T_2} - 300\frac{3k}{4e}T_1\frac{\log T_1/T_2}{T_1-T_2} - 300\frac{3k}{4e}\log T_2 = 300\frac{k}{e}\left(\log\frac{n'_F}{A'}-\frac{3}{4}\right)+\frac{V_\beta}{2}\left(\frac{\log T_1/T_2}{T_1-T_2}\right), \tag{21}$$

where V_β is again the activation potential in volts, corresponding to V_α in (6B). Its value can be deduced from the line obtained by plotting the left-hand side of (21), of which all quantities are known in terms of experimental results, against the usual $(\log T_1/T_2)/(T_1-T_2)$. Equation (21) should represent a better approximation to the facts than (6B), subject to the main assumption being valid, namely, that the semi-conductor is homogeneous.

3. *Thermo-Electric Effects in Heterogeneous Semi-Conductors*

The semi-conductors normally available for investigation are often not entirely homogeneous, and it is important to examine the anomalies that can occur as a result of heterogeneous structure. Two important types of heterogeneity can be envisaged. Thus, a semi-conductor may have a regular grain structure, but the concentra-

* It was thought conceivable that thermal expansion might have a noticeable effect on the Thomson coefficient, in so far as it discourages electron diffusion away from the hot regions. The approximate calculation of this effect can proceed by the same method as before, taking into account that the electron concentration at any temperature T must be reduced by a factor $1/(1+cT)$, where c is the (volume) coefficient of thermal expansion. Using this correction, it follows that $\sigma_s = \epsilon/2eT+(k/e)[3/4-cT/(1+cT)]$. Taking c as 10^{-5} centimetres per degree centigrade and T as 600 degrees absolute, substitution shows that the reduction of the Thomson coefficient by thermal expansion is negligible at all temperatures here considered.

tion of impurity centres may change systematically from one side of the solid to the other. Alternatively, the substance may be a conglomerate of grains, which differ individually in impurity content in a random manner, so that numerous large concentration gradients are distributed throughout the solid. In practice, the two cases may, of course, arise simultaneously.

Consider the effect of continuous concentration gradients on the Thomson coefficient of the semiconductor. To substitute for n in (12), when N is a function of x, we must put

$$dn/dx=(\partial n/\partial T)(dT/dx)+(\partial n/\partial N)(dN/dx), \quad (22)$$

which gives

$$\frac{dn}{dx}=n_x\left(\frac{\epsilon}{2kT_x^2}+\frac{3}{4T_x}\right)\frac{dT}{dx}+\frac{n_x}{2N_x}\cdot\frac{dN}{dx} \quad (23)$$

and, after substitution into (14), we obtain

$$\left(\frac{dE}{dT}\right)_x=\sigma_x=\left(\frac{\epsilon}{2eT_x}+\frac{3k}{4e}\right)+\frac{kT_x}{2eN_x}\frac{(dN/dx)}{(dT/dx)}. \quad (24)$$

It will be noted that this reduces to the previous form (16) when dN/dx is zero. Otherwise, the local value of the Thomson coefficient depends on the ratio of the concentration gradient to the temperature gradient.

If the specimen is entirely at a uniform temperature T, then the first term in (24) cannot make any contribution to the voltage, since its temperature integral between identical limits is zero. The second term, however, gives rise to an electromotive force E_0 between any two points that are characterised by impurity concentrations N_1 and N_2 respectively given by

$$E_0=(kT/2e)\int_{N_2}^{N_1}(1/N)dN$$
$$=(kT/2e)\log N_1/N_2. \quad (25)$$

This electromotive force may be termed a "distributed contact potential." The influence of this distributed contact potential on the conventional representation of energy levels is shown in Fig. 2. The voltage E_0 is not directly measurable by means of contact electrodes, because it is exactly balanced by equal and opposite contact potential at the two junctions.

If the junctions at the ends of a heterogeneous semi-conductor are *at the same temperature*, but temperature gradients exist elsewhere between the two contacts, then the contact potentials at the junctions and the distributed contact potential within the solid are not necessarily equal and opposite, and a residual electromotive force E_r can, therefore, appear between the electrodes, given by

$$E_r=\frac{k}{2e}\oint\frac{T_x}{N_x}\frac{(dN/dx)}{(dT/dx)}dT. \quad (26)$$

If dN/dx, as well as dT/dx, vary symmetrically between the two electrodes, E_r is zero, but if

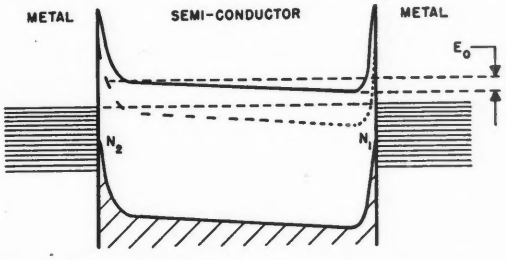


Fig. 2—Energy relations in an excess semi-conductor of non-uniform impurity content.

either variable is unsymmetrical, E_r will have a finite value, which may be of either sign.

If the junctions at the ends of the semiconductor are *not* at the same temperature, the normal thermo-electromotive force will, of course, appear in addition to the "distributed thermo-electromotive force" of magnitude

$$E_D=\frac{k}{2e}\int_{T_2}^{T_1}\frac{T_x}{N_x}\frac{(dN/dx)}{(dT/dx)}dT. \quad (27)$$

This electromotive force will have the greatest influence on the effective thermo-power when the temperature differences are small, and its presence should cause the apparent thermo-power to tend to infinity as $|T_1-T_2|\to 0$. When the normal thermo-electromotive force is small, the sign of the effective thermo-power may be entirely dominated by E_D.

In the general case, the equation for the total thermo-electromotive force developed between two points at temperatures T_1 and T_2 and characterized by constants A_1' and A_2' becomes too complicated to be of practical use, namely,

$$E=(kT_1/e)\log(n_F'/A_1')T_1^{\frac{3}{2}}$$
$$-(kT_2/e)\log(n_F'/A_2')T_2^{\frac{3}{2}}$$
$$-(3k/4e)(T_1-T_2)$$
$$+(\epsilon/2e)\log T_1/T_2+E_D. \quad (28)$$

If $A_1' \simeq A_2'$ and $|T_1 - T_2| \gg 0$, (21) should still be approximately obeyed.

These considerations show that as far as significant results can be obtained at all by experiments on heterogeneous materials, they necessitate the use of large temperature differences to minimize the effect of the distributed thermo-electromotive forces. In general, the properties of semi-conductors are very sensitive to changes of condition during their preparation. The existence of small variations of impurity content is, therefore, to be expected in sintered semi-conductors (e.g., as a result of uneven temperature distribution during their preparation) and probably in most others, particularly if the specimens are of considerable size.

The thermo-electric anomalies caused by the presence of concentration gradients are still expected to appear if the material concerned is a heterogeneous conglomerate of the second type, as envisaged above, provided that differences exist between the *average* impurity content of various regions in the solid.

4. Preparation and Micro-Structure of Titanium Dioxide Specimens

The specimens for the present investigation were prepared by pressing fine titanium dioxide powder into rectangular rods 10 centimetres long and 1 centimetre square under standard conditions. These rods were then pre-fired in air at about 1050 degrees centigrade for 1.5 hours. During this process, some volatile impurities are eliminated and the powder gains cohesion. The specimens were then slowly inserted into an electrically heated tube-furnace which was already at a temperature of 1000 degrees centigrade. During the four hours following, the temperature was raised to 1500 degrees and a mixture of 70 per cent nitrogen and 30 per cent hydrogen was passed through the tube. The specimens remained at the high temperature for four hours before the oven was allowed to cool. Each batch of four samples was then divided into two pairs. One pair was withdrawn from the oven after a cooling period of 25 minutes (1200 degrees) and then cooled further in the open air. These specimens will be referred to as "chilled," to distinguish them from those that were more "slowly cooled," i.e., that remained in the oven until its temperature had fallen to 700 degrees after 1.5 hours. These specimens were then further cooled while surrounded with asbestos packing. Owing to differential contraction, the number of breakages during "chilling" was considerable.

Chilled and slowly cooled specimens were expected to differ with regard to the concentration of ions which are "frozen" in interstitial positions, but no information was available as to whether the above differences in the rates of cooling are sufficient to have any appreciable effect in this respect.

The high-temperature sintering process in a reducing atmosphere results in a further elimination of volatile impurities and in a volume shrinkage of between 44 and 56 per cent. The reduced specimens are greyish-blue, with a metallic lustre, very hard and brittle. Their density is approximately 3.8 grams per cubic centimetre and does not vary between wide limits. The specimens are only slightly porous. X-ray powder-photographs have shown that they consist of rutile, and there is no evidence of the existence of any second crystalline phase.

The specimens were examined under the microscope in the form of very thin layers (20–30 microns) and as a very fine powder (capable of passing through a No. 100 mesh), both mounted in Canada balsam. It was found that these sintered bodies are not of homogeneous composition, but consist of four types of grains.

A. Small, faintly yellow grains, slightly bi-refringent and slightly pleocroic, having the characteristic appearance of normal rutile.

B. Grains of similar appearance except as regards their colour, which varies from slightly grey to dark blue-grey.

C. Fairly large black grains, often needle shaped, interspersed amongst a continuum of types *A* and *B*, and much more frequent in regions where the latter type is common.

D. Foreign impurities of two kinds: small grains of a red crystalline material, thought to be haematite, and somewhat larger grains of a material whose refractive index is very nearly the same as that of Canada balsam, probably quartz. Both kinds are only present in minute proportions and could not make any measurable contribution to the conductivity.

It was later found that the concentration of black needle-shaped crystals is greatest in the chilled specimens and smallest in specimens of low conductivity. The grain structure of the

specimens is usually regular throughout the bulk of the material, but near the surface the proportion of yellow rutile grains increases sharply. This surface layer, which is of much higher resistivity than the rest of the material, is best developed on the slowly cooled specimens, i.e., those which have been for a longer time under conditions favourable for re-oxidation. The photo-micrograph, Fig. 3, shows the boundary

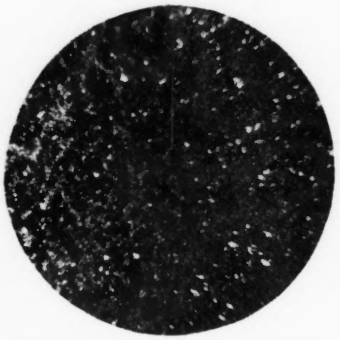


Fig. 3—Photo-micrograph of titanium dioxide specimen approximately 1.4 millimetres in diameter. Bulk material is at the left and the surface layer at the right.

between the bulk of the material on the left and the surface layer on the right. When a specimen is broken up into a fine powder, there is some tendency for grains of uniform colour (and hence, presumably, of uniform composition) to remain intact.

The existence of this grain structure and the correlation of some of its characteristics with the method of preparation and the resistivity of the samples must inevitably be of great importance for the interpretation of the experimental results. There is reason to believe that these considerations are not confined to the present case, but are probably relevant to many other semi-conductors and to many previous measurements. It is important to realise that the normal theory of semi-conductors is not quantitatively applicable to conglomerates of the type here described.

5. Experimental Technique

As far as thermo-electric measurements on semi-conductors are concerned, either of two distinct methods can be adopted. One method consists of establishing a large temperature difference between two metallic junctions on the semi-conductor and measuring the resulting (total) thermo-electromotive force. The other consists of raising (or lowering) the whole specimen to the desired average temperature, and measuring the thermo-electromotive force due to a small temperature difference between the two junctions. If this temperature difference is sufficiently small, the measurement yields the true thermo-power at the temperature concerned. This is the method that has usually been adopted in the past, but it suffers from the grave draw-backs already discussed in Section 3. This is thought to be the reason for the low accuracy claimed by previous experimentors. Thus Morton's results[15] are consistent only within ± 5 per cent, and those of Vogt[13] within even wider limits. On these grounds and also for reasons of greater experimental convenience, the first method was used during this investigation.

The simple arrangement is shown schematically in Fig. 4. The semi-conducting rod under test rested on two supports and was insulated from them by thin sheets of mica. On each side, a thin brass disc was firmly clamped to the specimen. The pressure was adjusted until the

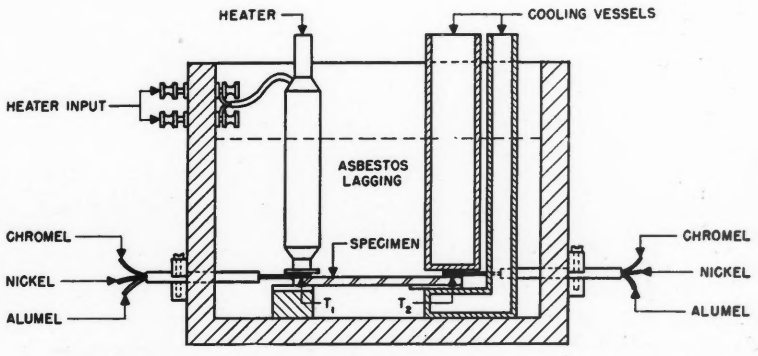


Fig. 4—Experimental arrangement for thermo-electric measurements.

contact resistance remained approximately constant. These brass discs, of which Fig. 5 is an enlarged representation, served as electrodes. Their tempertures could be determined by means

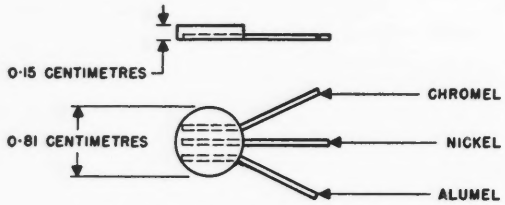


Fig. 5—Thermo-couple electrodes.

of the attached "Chromel-Alumel" thermo-couples. The thermo-electromotive forces otherwise measured are those between the semi-conductor and nickel. One of the discs was heated by means of the heater, which was operated from a variable auto-transformer. To protect the associated galvanometer circuits in the event of a breakdown, the heater was electrically insulated from the electrode by a thin sheet of mica. The other electrode could be cooled from below and above by filling the vessels which form the clamp with solid carbon dioxide or liquid air. Cooling from below and above was found to be essential to make the best possible use of the cooling agent. The specimen was surrounded with asbestos lagging so as to exclude draughts, which were otherwise found to have a deleterious effect on the temperature stability of the system. To test whether slight oxidation of the electrodes could introduce substantial errors, some of the measurements were repeated, using gold-plated electrodes. Consistent results were obtained, and it can thus be assumed that if any oxide film is formed on the electrodes at temperatures below 450 degrees centigrade, it is too thin to introduce any appreciable temperature gradients.

The whole arrangement is associated with a suitable potentiometer circuit. Each complete measurement thus demands the determination of T_1 and T_2 (by reference to oil-immersed cold junctions at a known temperature) and E, the resulting thermo-electromotive force between the semi-conductor and nickel.

The resistivity of the same specimens was measured between room temperature and 200 degrees centigrade, using the simple jig illustrated in Fig. 6. The potential probes, which were held in position by springs, rested on thin graphite lines which served to reduce the resistance in the probe circuit. The jig was mounted in an oven in turbulent air, and the potential difference between the probes measured at various temperatures for a constant current of 10 milliamperes.

It soon became evident that special precautions must be taken to eliminate unwanted thermo-electromotive forces in the potential probe circuit, or to correct for their presence. In view of the large thermo-powers developed by the junctions of the semi-conductor and the probes, very small temperature gradients along the specimen can lead to substantial errors in the determined resistivities. As small temperature gradients could not be entirely eliminated, a correction was made by measuring the thermo-electromotive force between the potential probes when the specimen was disconnected from the battery circuit. This procedure greatly improved the consistency of the results.

6. Resistivity Measurements

The room-temperature (20 degrees centigrade) resistivity of the present specimens lies between 0.10 and 0.18 ohm–centimetres and is independent of the applied voltage. The temperature coefficient at 20 degrees centigrade is of the order of 1 per cent per degree centigrade and decreases with increasing temperature. All resistivity measurements between room temperature and 200 degrees centigrade were found to be strictly repeatable. Fig. 7 shows that—as might be expected—there is some correlation between the resistivity at room temperature and the percentage shrinkage during preparation. (All resistivities quoted are "effective" values. The

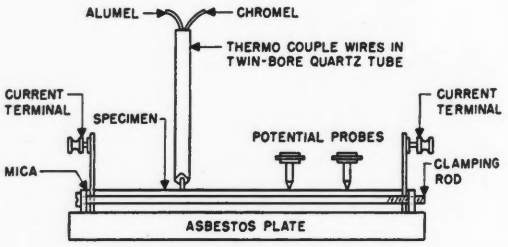


Fig. 6—Experimental arrangement for resistivity measurements.

existence of the more highly resistive surface layer of the specimens was neglected in the calculation.)

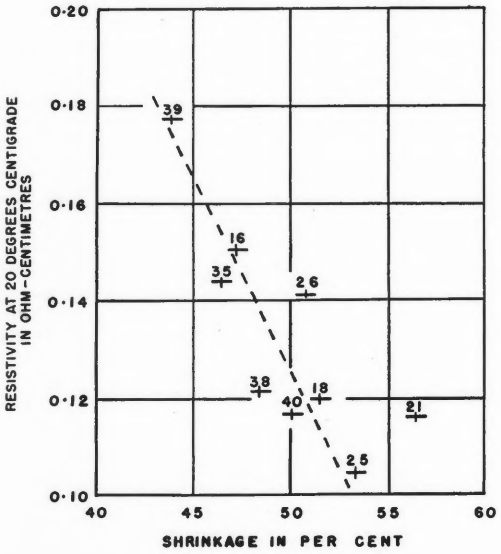


Fig. 7—Relation between resistivity at room temperature and shrinkage of specimens during firing. The specimen numbers are quoted.

No significant difference between chilled and slowly cooled specimens was noted. This seems to indicate that the process of chilling was not sufficiently drastic to make any appreciable difference in the concentration of ions frozen in interstitial positions. Alternatively, it may be that the critical temperature below which the presence of interstitial ions affects the resistivity measurements is considerably lower than room-temperature. (See page 160 of reference 24.)

To obtain the effective activation energies of the specimens, the results can be plotted in two alternative ways. For small temperature ranges, the expression

$$R = R_0 \exp [eV_a/2kT] \qquad (29)$$

is often a sufficiently accurate representation of the experimental results. A value for V_a can then be deduced from the slope of the line obtained by plotting R against $1/T$. However, for measurements over the temperature range here employed, these lines are all noticeably curved, so that only average gradients can be found from them. This means that the temperature dependence of R_0 is not negligible compared with that of the exponential term. This does not seem to have been observed during some previous experiments on titanium dioxide.[25] An expression for R, which is theoretically more satisfactory than (29), is

$$R = KT^{\frac{1}{2}} \exp [eV_b/2kT], \qquad (30)$$

where K is a constant. A straight line should thus be obtained by plotting $\log R/T^{\frac{1}{2}}$ against $1/T$. This was found to be the case, as shown in Fig. 8 for three typical specimens.†

A survey and summary of the results will be found in Table I. Wherever applicable, intercepts have been corrected for the different cross-sectional areas of the various specimens. It will be noted that V_a and V_b, the two activation potentials deduced from (29) and (30) respectively are not identical. V_b must be regarded as the more correct and significant value, even if its fundamental meaning is still somewhat in doubt.

† For convenience, a multiple of the voltage V_p between the potential probes for a constant current (10 milliamperes) is used instead of the resistance R.

TABLE I

Specimen Number	Type	Resistivity at 20°C Ohm-cm	10^3K Ohm-cm	V_a Volts	V_b Volts	V_α Volts	V_β Volts	$\frac{3k}{e}\log\frac{n_F}{A}\times10^5$ $mV/°C$
			(30)	(29)	(30)	(8)	(21)	(8)
16	chilled	0.151	0.179	0.081	0.127	0.033	0.076	0.248
18	chilled	0.120	0.114	0.088	0.139	0.060	0.116	0.165
21	chilled	0.116	0.082	0.078	0.127	0.060	0.122	0.193
25	slow cooled	0.105	0.133	0.076	0.129	0.046	0.104	0.191
26	slow cooled	0.141	0.176	0.072	0.125	0.037	0.089	0.207
35	slow cooled	0.144	0.202	0.071	0.117	0.036	0.088	0.275
38	slow cooled	0.122	0.073	0.118	0.168	0.014	0.076	0.134
39	slow cooled	0.177	0.123	0.107	0.152	0.025	0.084	0.183
40	slow cooled	0.117	0.077	0.108	0.166	0.029	0.084	0.162

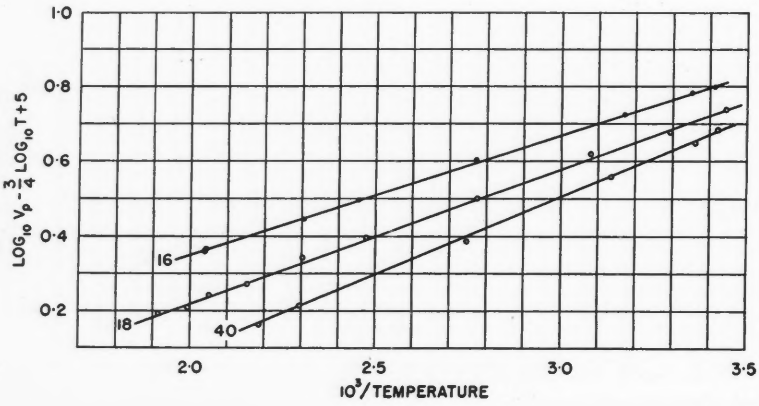


Fig. 8—Effect of temperature on the resistance of three specimens. $R = KT^{\frac{3}{4}} \exp [eV_b/2kT]$. V_p is the probe voltage and is proportional to R.

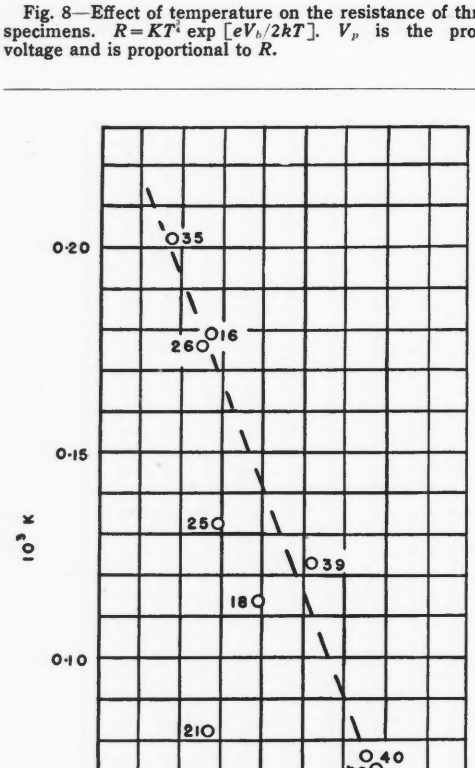


Fig. 9—Correlation between experimental values of concentration constant K and effective activation energy (V_b).

Meyer and Neldel[26] have proved the existence of a definite relation between V_a and R_0, contrary to earlier theoretical expectations. Fig. 9 shows that some such correlation also exists in the present case. (V_b and K are actually plotted.) The values obtained for V_a fall within narrow limits compared with those quoted by the above authors. (From 0.049 to 1.044 volts.)

7. *Thermo-Electric Measurements*

Measurements of the thermo-electromotive force of titanium dioxide–nickel couples were carried out over a temperature range from −176 to +400 degrees centigrade. Fig. 10 gives the results of one such complete series, plotted in accordance with (21). The temperatures T_1 and T_2 are marked for representative points. It will be seen that over this wide temperature range the results are substantially in agreement with theory.‡ In general, the larger the temperature difference between the two junctions, the closer is the agreement of the results with the predicted straight-line characteristic. The deviations which exist appear to be typical for the present specimens. Large temperature gradients in the neighbourhood of the cold junction result in slightly low values for the average thermo-power, whereas large gradients in the neighbourhood of the hot junction do the reverse. It is to be expected that the latter occur shortly after the heater current is switched on, the former mostly during the initial period of cooling by means of solid carbon dioxide or liquid air. As the cooling and heating processes continue, the temperature gradients become less sharp and, if $T_1 - T_2$ is sufficiently large, their effect becomes negligible.

None of the points in Fig. 10 corresponds to a temperature difference of less than 70 degrees centigrade. If $T_1 - T_2$ is appreciably smaller, large deviations from the predicted behaviour

‡ When the experimental results are plotted in accordance with (8), they likewise give a straight line, though of smaller slope.

are observed. These deviations are interpreted as the direct consequence of the distributed thermo-electromotive forces that are expected to arise from the heterogeneous structure of the present specimens. The region of small temperature differences was further investigated in an attempt to isolate the normal thermo-electromotive force from the distributed thermo-electromotive force. To do this, a small asymmetrically wound heating coil was placed between the two main junctions (in the neighbourhood of the cold electrode) to produce the desired asymmetrical temperature distribution. Due to the presence of this subsidiary heater, T_2 rose to a constant value. The main heater was then switched on, and T_1 raised to and above T_2. The resulting thermo-electromotive forces were measured, and the results plotted in Fig. 11. They show clearly that the apparent average thermo-power tends to infinity as $|T_1-T_2|$ tends to zero, which means that a finite thermo-electromotive force E_r exists even when $T_1=T_2$. It will also be seen that the thermo-power of the couple is always positive provided $|T_1-T_2|$ is large, but can be of either sign when $|T_1-T_2|$ is small. If the subsidiary heater is displaced, the relative positions of the temperature and concentration gradients are also changed and this can be shown to have a considerable effect on the magnitude of E_r. The pairs of lines marked a, b, and c on Fig. 11 refer to three positions of the subsidiary heater. It will be noted that E_r changes sign between positions a and b, which seems to indicate that the average impurity content varies approximately symmetrically along the sample. Under the conditions to which Fig. 11 refers, E_r is $+5.9$ millivolts for position a, $+1.5$ millivolts for position b, and -6.4 millivolts for position c. There is, of course, some position between a and b for which E_r is zero.

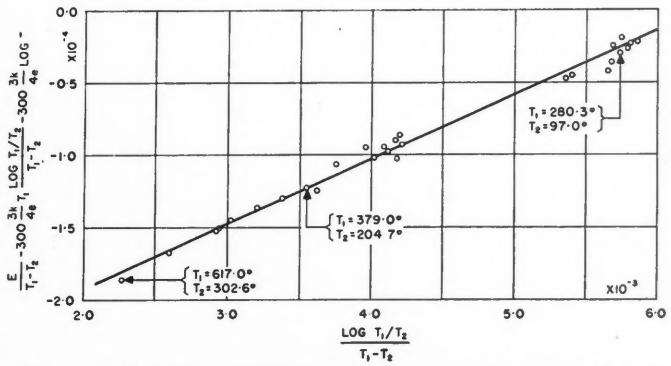


Fig. 10—Graphical determination of activation energy from thermo-electric measurements. Plot of experimental results in accordance with (21). Temperatures T_1 and T_2 are indicated for representative points. Specimen 26.

Thermo-electromotive forces of this type have already been observed in the case of metals by

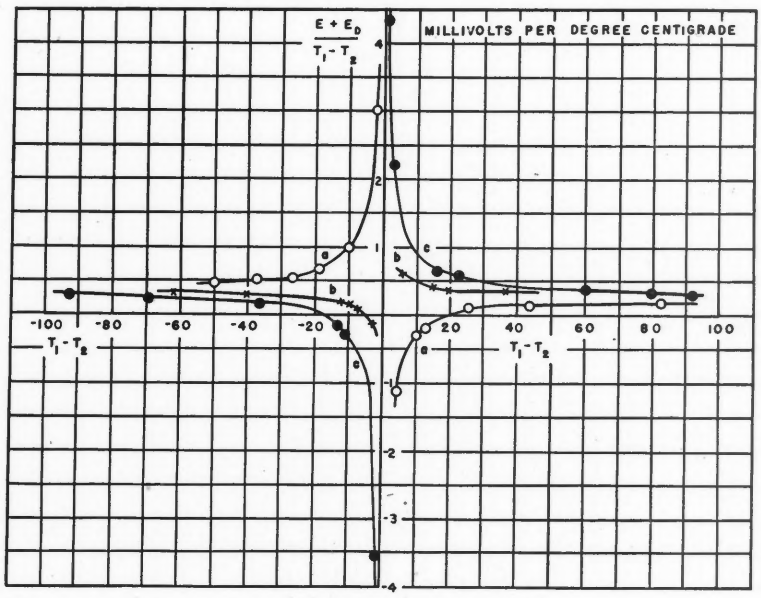


Fig. 11—Effect of distributed thermo-electromotive forces on apparent average thermo-power of the heterogeneous specimen 36.

Feng and Band,[27] though these authors did not associate them with any heterogeneous structure of the materials used. The distributed thermo-electromotive forces observed on the present samples are higher than those of metals by about two orders of magnitude.

The sign of the thermo-power of a semi-conductor–metal couple is generally taken as evidence for the character of the semi-conductor.[19] The demonstrated existence of distributed thermo-electromotive forces means, however, that this criterion is not valid unless the semi-conductor is known to be entirely homogeneous or unless the measurements are carried out for large temperature differences. In the general case, there is a further ambiguity owing to the fact that $\log n_F/A$ in (7) can be of either sign. Hence, it is important to know not only the sign of the thermo-power at a given temperature, but the exact characteristics of the couple over the widest possible temperature range. Conclusions derived from single measurements using small temperature differences can be entirely misleading in this respect.

Values for the activation potentials V_α and V_β were deduced from the experimental results and incorporated in Table I. It will be noted that V_b and V_β are not identical, but agree as regards general order of magnitude. Curiously enough, the agreement is best in the cases of specimens 21, 18, and 25, i.e., those which on Fig. 9 differ most prominently from the remaining samples. Otherwise, we have $V_\alpha < V_a$ and $V_\beta < V_b$, without any significant correlation between these quantities.

8. Discussion

Two important problems arise in connection with the present—and indeed many previous—measurements. Firstly, it will be observed that the nine specimens on which measurements have been carried out have yielded 15 different values for the activation energy. There is, in fact, nothing to suggest that V_β and V_b cannot vary continuously over a wide range of voltages. If eV_b (or eV_β) is the width of the energy gap between the impurity level and the conduction band of the solid, these results would mean that the width of this gap can vary continuously. This is very hard to envisage. On the contrary, it is to be expected that there are only a limited number of positions in the unit cell of the crystal lattice in which an interstitial atom can be accommodated, and while each of these may be associated with a different activation energy, the total number of such values should be small. Secondly, there is no *a priori* reason for any correlation between the activation energies and the impurity contents of various specimens, unless, of course, the impurity content is so large that the centres interfere with one another. This interference, although possible, does not seem to be the general solution of the problem, since Meyer[25] has shown the existence of this correlation even for specimens which were excellent insulators at room temperature, i.e., specimens in which the concentration of impurity centres must have been very small indeed.

Since the present specimens are not homogeneous, but contain grains of different impurity content, it appears reasonable to associate the multiplicity of values for V_a and V_b (similarly V_α and V_β) with the multiplicity of proportions in which grains of different purity may constitute any particular specimen. V_a and V_b then have the significance of "effective" activation energies, and are dependent on the activation energies of the different types of grains. The problem in its simplest form can be formulated as follows: What are the effective values of R_0 and V_a of a solid conglomerate composed of semi-conducting grains whose individual constants or group constants are $r_1, r_2, \ldots, r_n$, and $V_{a1}, V_{a2}, \ldots, V_{an}$ (so that $R_n = r_n \exp[eV_{an}/2kT]$), and which are present in concentrations $c_1, c_2, \ldots, c_n$, respectively? It is assumed that r_n depends only on the concentration of impurity centres within each type of grain and can, therefore, vary continuously, whereas the activation energies V_{an} are limited to a small number of definite values. To obtain an approximate solution, Borelius and co-workers[28] have suggested the application of a much earlier result due to Lichtenecker[29] on the resistivity of mixtures. Lichtenecker's calculations are mainly intended for two-dimensional models of two almost continuous constituents, but could be generalized in the following form.

The resistivity of the aggregate must be given by some function

$$R=F(R_1, R_2, \ldots., R_n;\quad c_1, c_2, \ldots., c_n). \quad (31)$$

Lichtenecker also imposes the (arbitrary) condition:

$$1/R=F(1/R_1, 1/R_2, \ldots., 1/R_n;\quad c_1, c_2, \ldots., c_n). \quad (32)$$

Further, we must have $R=R_n$ if $c_n=1$ and $R=0$ if $R_n=0$ and $c_n\neq 0$. The simplest expression which satisfies these conditions is

$$R=\prod R_n^{c_n g_n}, \quad (33)$$

where the quantities g are parameters that depend on the geometrical configuration of the grains. For dimensional reasons, we must also have $\Sigma c_n g_n=1$.

Equation (33) cannot, of course, represent more than a first approximation. The treatment does not take account of the potential barriers between individual grains, nor does it seem probable that the characteristics of the configuration can be satisfactorily expressed by a single parameter for each constituent. Subject to these reservations, however, the expression is partially successful, in so far as it predicts the existence of an effective activation energy eV_a, where

$$V_a=\Sigma c_n g_n V_{an}, \quad (34)$$

a quantity which must be expected to vary from sample to sample. Also, since

$$R_0=\prod r_n^{c_n g_n}, \quad (35)$$

some correlation between the values of V_a and R_0 is to be expected for specimens which have similar values of g.

If the resistivity of the specimens were entirely due to the properties of individual grains and not to structural factors, a close negative correlation should exist between resistivity and the concentration of impurity centres. If the present considerations are correct, the intercepts of the lines obtained by plotting the results of thermo-electric measurements in accordance with (8) should arrange the specimens approximately in the reversed order of their average impurity content. Fig. 12 shows that a correlation between these intercepts and those on Fig. 8 (representing the resistivities) does indeed exist.

The fact that the correlation is not complete can be interpreted as meaning that the average impurity content of the grains and their geometrical configuration make comparable contributions to the resistivity.

The precise manner in which the different types of grain affect the resistivity of the substance as a whole remains a subject for future investigation. Because the grains presumably differ as regards their impurity content, the Schottky barriers at each grain contact represent an additional source of resistance. Each barrier of this type would normally have non-linear and even slightly rectifying conduction characteristics, if the applied voltage is large. For small voltages, these resistances should be ohmic. The resistivity of the specimens as a whole was found to be strictly independent of the applied voltage, and it must therefore be assumed that the voltage developed across individual barriers is too small to produce non-linear conduction, but that the total voltage across all barriers is an appreciable fraction of the total voltage applied.

The correlation of results shown in Fig. 12 confirms that semi-conductor–metal couples of high thermo-power are also expected to have a high resistance. This provides a natural limit of practical applicability, because high thermo-

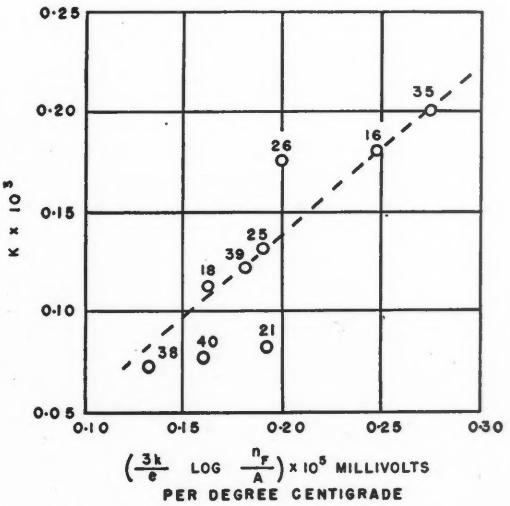


Fig. 12—Correlation between thermo-electric and resistivity measurements.

powers can only be usefully employed as long as the internal resistance of the couples is reasonably low.

With regard to the experimental values for the activation potential, agreement between V_b and V_β can only be expected as regards order of magnitude in the case of specimens that are not homogeneous. The internal mechanism by which activation energies can be "compounded" from a series of different contributions is still practically unexplored. There is no *a priori* reason to assume that the resultant value should be identical for conductive and thermo-electric processes. In the latter case, the heterogeneous structure is expected to result in "internal short-circuits" which would tend to diminish the total thermo-electromotive forces. This may account for the fact that V_b is always greater than V_β, but the problem does not seem to lend itself to more precise analysis. For homogeneous materials V_b and V_β should, however, be identical, and it would be of great interest to repeat the present experiments on substances that can be obtained in the form of single homogeneous crystals of suitable size.

9. *Summary of Main Conclusions*

A. A simple theory of thermo-electric effects, based on the electronic energy relations at a contact, leads to an expression (21) that is in qualitative agreement with experimental results over the temperature range −176 to +400 degrees centigrade.

B. The Thomson coefficient of a semi-conductor can be deduced from the equilibrium condition between thermal diffusion and electric fields. (Equations (16) and (24).)

C. Sintered titanium dioxide specimens consist of grains of different colour and hence, presumably, of different impurity content.

D. In a semi-conductor that is not homogeneous (i.e., in which the average concentration of impurity centres is subject to local variations), "distributed contact potentials" must exist throughout the solid, even when the temperature is uniform. Distributed contact potentials cannot be directly measured by means of contact electrodes, since they are equal and opposite to the normal contact potentials.

E. If temperature gradients exist in a heterogeneous semi-conductor, "distributed thermo-electromotive forces" appear in addition to the normal thermo-electromotive force. As a result, the apparent average thermo-power $(E_D+E)/(T_1-T_2)$ tends to infinity as the temperature difference between the junctions tends to zero.

F. Thermo-electric measurements can be used to grade a series of specimens in the order of their average impurity content. Comparison with the results of resistivity measurements then gives some guidance as to the extent to which the grain configuration affects the resistivity. The influence of configuration is detectable for the present specimens.

G. If only one mechanism of conduction is active, thermo-electric measurements can be used to determine the character of a semi-conductor (excess of deficit type), provided that large temperature differences are used and certain precautions observed.

H. The magnitude of the distributed thermo-electromotive forces gives some guidance on the homogeneity of specimens.

I. A correlation exists between the room-temperature conductivity of the specimens and the percentage shrinkage during preparation. The resistivity of the specimens is independent of the applied voltage.

J. Resistivity measurements between room temperature and +200 degrees centigrade are strictly repeatable. No differences have been observed between chilled and slowly cooled specimens.

K. In general, semi-conductors of high thermo-power have a high resistivity, and vice versa.

L. The temperature dependence of the resistivity between room temperature and 200 degrees centigrade is well expressed by the theoretical equation $R=KT^{\frac{1}{2}} \exp[eV_b/2kT]$. In the case of a heterogeneous solid, the interpretation of the energy term eV_b is still in some doubt. It is thought probable that eV_b represents some form of effective activation energy, e.g., of the kind arising from Lichtenecker's equations.[29]

M. Rough agreement can be reached between the activation energies deduced from resistivity and thermo-electric measurements respectively. Exact agreement cannot be expected when the specimens are heterogeneous.

10. *Acknowledgments*

The author wishes to make grateful acknowledgment to Prof. R. W. Ditchburn for placing research facilities at his disposal and for the stimulating direction of this work, to Standard Telecommunication Laboratories Limited for a research grant, to Mr. T. R. Scott (of that organization) for his personal interest in this research and for kindly supplying experimental specimens, to the Director, Royal Aircraft Establishment for the loan of certain test equipments, to Dr. T. B. Rymer for many helpful discussions and much valuable advice, and to Mrs. C. G. Stephenson for assistance with the measurements and calculations.

11. Bibliography

[1] Earle, "Electrical Conductivity of Titanium Dioxide," *Physical Review*, v. 61, p. 56; 1942.

[2] Busch, "The Electrical Conductivity of Silicon Carbide," *Helvetica Physica Acta*, v. 19, p. 463; 1946.

[3] Busch and Labhart, "On the Mechanism of the Electrical Conductivity in Silicon Carbide," *Helvetica Physica Acta*, v. 19, p. 167; 1946.

[4] Mönch, "Theory of Thermo-Emfs Between Semi-Conductor and Metal, Based on Fermi Statistics," *Zeitschrift für Physik*, v. 83, p. 247; 1933.

[5] Mönch and Stechhofer, "Thermo and Contact Emfs on Cu_2O," *Zeitschrift für Physik*, v. 84, p. 59; 1933.

[6] Mönch, "Thermo and Contact Emfs on Cuprous Oxide," *Naturwissenschaften*, v. 21, p. 367; 1933.

[7] Mönch, "On the Temperature Dependence of Contact Emfs on Cu_2O," *Zeitschrift für Physik*, v. 90, p. 576; 1934.

[8] Mönch, "Thermo-Voltage, Peltier Heat and Photo Effect on the Element $Cu-Cu_2O-Cu$," *Zeitschrift für technische Physik*, v. 16, p. 361; 1935.

[9,10] Mönch, "Thermo-Emfs on the Element Metal Semi-Conductor Metal," *Annalen der Physik*, v. 26, p. 481; 1936: and v. 34, p. 265; 1939.

[11] Schweickert, "Thermo-Emfs on the Element Metal Semi-Conductor Metal II," *Annalen der Physik*, v. 34, p. 250; 1939.

[12] Rohde, "Thermo-Emfs on the Element Metal Semi-Conductor Metal III," *Annalen der Physik*, v. 34, p. 259; 1939.

[13] Vogt, "Electrical Measurements on Cuprous Oxide," *Annalen der Physik*, v. 7, p. 183; 1930.

[14] Borelius and Gullberg, "Thermo-Electric Power of Liquid and Solidified Selenium," *Arkiv för Matematik Astronomi och Fysik*, v. 31A, n. 17; 1945.

[15] Morton, "Semi-Conducting Properties of Lead Sulphide," *Transactions of the Faraday Society*, v. 43, p. 194; 1947.

[16] Anderson and Morton, "Semi-Conducting Properties of Stannous Sulphide," *Transactions of the Faraday Society*, v. 43, p. 185; 1947.

[17] Johnson and Lark-Horovitz, "Theory of Thermo Electric Power in Germanium," *Physical Review*, v. 69, p. 259; 1946.

[18] Lark-Horovitz, Middleton, Miller, Scanlon, and Walerstein, "Electrical Properties of Germanium Alloys II, Thermo-Electric Power," *Physical Review*, v. 69, p. 259; 1946.

[19] Andrews, "Thermo-Electric Power of Cadmium Oxide," *Proceedings of the Physical Society*, v. 59, p. 990; 1947.

[20] Fischer, Lepsius, and Baerwindt, "On Silicon and its Position in the Thermo-Electric Series," *Physikalische Zeitschrift*, v. 14, p. 439; 1913.

[21] Baumbach and Wagner, "On the Electrical Conductivity of Zinc Oxide and Cadmium Oxide," *Zeitschrift für physikalische Chemie*, B, v. 22, p. 199; 1933.

[22] Fischer, Dehn, and Sustmann, "On the Thermo-Power of Multiple Oxide Mixtures," *Annalen der Physik*, v. 15, p. 109; 1932.

[23] Fowler, "Statistical Mechanics," Cambridge University Press, Cambridge, 1936; p. 427.

[24] Mott and Gurney, "Electronic Processes in Ionic Crystals," Clarendon Press, Oxford, 1942.

[25] Meyer, "Contribution on the Electrical Conductivity of Semi-Conducting Materials," *Zeitschrift für technische Physik*, v. 16, p. 355; 1935.

[26] Meyer and Neldel, "On the Activation Energy of Semi-Conductors," *Physikalische Zeitschrift*, v. 38, p. 1014; 1947.

[27] Feng and Band, "Longitudinal Thermo-Electric Effect : Copper," *Proceedings of the Physical Society*, v. 46, p. 515; 1934.

[28] Borelius, Pihlstrand, Anderson, and Gullberg, "Resistance of Liquid and Solidified Selenium," *Arkiv för Matematik Astronomi och Fysik*, v. 30A, n. 14; 1944.

[29] Lichtenecker, "Electrical Resistance of Artificial and Natural Aggregates," *Physikalische Zeitschrift*, v. 25, p. 169; 1924.

[30] Hogarth, "Variation of Thermo-Electric Power of Electronic Semi-Conductors with Vapour Pressure," *Nature*, v. 161, p. 66; 1948.

Relationship Between Rate of Transmission of Information, Frequency Bandwidth, and Signal-to-Noise Ratio*

By C. W. EARP

Standard Telephones and Cables, Limited, London, England

THE history of communication is reviewed to show the gradual growth of modern theory, stressing in particular the major steps made by Carson with his sideband theory, and by Armstrong with his demonstration of improved quality of communication by use of increased bandwidth.

Established principles of communication are considered to show how they may be fitted together to form a coherent theory. After estimating the exact benefit to be gained by the use of the right balance between rate of transmission of information and the frequency bandwidth used in established expanded-band systems of communication, consideration is given to what particular features of transmission systems appear to be necessary for best performance.

Arising out of the above, a comparatively unknown system of modulation, tentatively called "step modulation," is shown to have all known characteristics of efficiency. The exact performance of the system is examined, and it is found that the process of bandwidth expansion to yield improved demodulated signal-to-noise ratio may be inverted to provide communication through reduced bandwidth at the cost of signal-to-noise ratio.

Finally, based upon the above analysis, a hypothesis is put forward to the effect that the new system may define the theoretical limit of efficient use of frequency bandwidth for any transmission system, for the case when the only known characteristic of the information wave is the frequency band that contains it.

* A symposium on "Recent Advances in the Theory of Communication" was presented at the November 12, 1947, meeting of the New York Section of the Institute of Radio Engineers. Four papers were presented by A. G. Clavier, Federal Telecommunication Laboratories; B. D. Loughlin, Hazeltine Electronics Corporation; and J. R. Pierce and C. E. Shannon, both of Bell Telephone Laboratories. The paper by Mr. Clavier was entitled "Evaluation of Transmission Efficiency According to Hartley's Expression of Information Content" and has been submitted to the Institute for publication.

1. Object of Paper

It is the object of this paper to produce an integrated modern theory of communication. At the present time, there appears to be considerable confusion concerning what use can be made of frequency bandwidth; various established principles do not appear to be co-ordinated into one general theory. Such a co-ordination should, it is believed, not only do much to establish a useful comparison between the efficiencies of well-known expanded-band systems of transmission, but may even permit us to specify a new ideal or maximum possible efficiency for the transmission of information according to the frequency bandwidth available.

The analysis does not include the conception of coding of specialised forms of information, but comprises only the problem of transmission of *any* waveform. No advantage will, for example, be taken of the fact that human speech usually carries much more information than the words that it comprises: we are here only concerned with transmitting any waveform with the maximum possible fidelity. Though the possibility of bandwidth compression arises, for the purpose of this study this must not be achieved by any coding process that might convert a voice showing particular personal characteristics, dialect, or "feeling" into a "mechanical" voice.

2. Historical Review of Communication Theory

The earliest electrical signalling was, of course, in the form of a code, information being carried in the form of a keyed direct current. Multichannel operation was later achieved by the use of keyed alternating currents of different frequency, and at this stage problems of bandwidth were already becoming significant.

Electrical transmission of speech, which was made possible by translation of an acoustic wave by a microphone into a corresponding electric wave, marks an all-important advance in practical technique. It is interesting to note that before the days of radio, but following the use of equivalent telegraph systems, multiple speech transmission was visualised by the distributor or time-sharing method.

The modulation of a carrier wave by speech currents is probably the next major step, and this immediately brings us to the time when the communication world became conscious of a relationship between signal-to-noise ratio and selectivity. Sideband theory was not, of course, available; the necessary frequency bandwidth for the transmission of speech on a carrier wave was not appreciated, though strangely enough, the bandwidth requirements of telephone lines for transmission of the speech wave itself were well understood.

The development of sideband theory, and Carson's conception of the single-sideband method of transmission probably represent the greatest contribution to communication theory up to the present. The single-sideband method persisted for a very long time as the fundamental ideal of transmission, and to this day it represents the best standard of efficiency for making comparison between other systems.

Both before and after the conception of single-sideband transmission, a great deal of work was carried out on other lines in an attempt to suppress noise. All the methods developed, however, were based either on a fallacy or on special conditions such as the presence of a particular type of noise. For example, the use of a wide receiving bandwidth for amplitude-modulated transmission and the use of a limiter to remove very large but short peaks of noise are quite effective, and may give a greater signal-to-noise ratio than the use of the same power in single-sideband transmission. Such systems, however, as Carson himself pointed out, involve particular assumptions with regard to the prevailing noise.

The Carson theory that, when noise is completely random in nature, the signal-to-noise ratio of a transmission system depends entirely upon the power developed in the signal sidebands, held good for many years. In justice to Carson, it should be added that the theory still holds good in cases where final signal-to-noise ratio is rather low. When noise power, receivable in a bandwidth equal to the information-wave bandwidth, is equal to or greater than the receivable signal power, then single-sideband transmission stands as an ideal that cannot be surpassed in signal-to-noise performance by any system. In other words, if only ability to communicate is considered, then single-sideband working represents a standard of perfection; it is only high-quality communication that has caused a revision of fundamental theory or, rather, opened up a new field of possibility.

Up to the date of Carson's teaching, improvement in signal-to-noise performance was attempted by reduction of transmission bandwidth. Having been set a very convincing lower limit for bandwidth, no revolutionary improvement now appeared possible. Probably, the natural attempts to minimise bandwidth actually retarded the communication world in its next major step.

Armstrong appears to be the first to have established that in conditions of small or moderate, but not very great, noise of a completely random nature, the signal-to-noise performance of single-sideband transmission can be radically improved by the use of greater bandwidth. His convincing demonstration of the new standard of quality of transmission, which can be achieved by broad-band frequency modulation, represents an important advance in communication technique.

At about the same time, A. H. Reeves was apparently conscious of the same principle in theoretical studies of pulse-time modulation, but practical development was not so rapid as in the case of frequency modulation.

Since the outbreak of war in 1939, little has been written of communication theory. Doubtless the principles established by Carson and Armstrong are now well co-ordinated in the minds of most communication engineers, but the present paper attempts to establish clearly, and without appreciable mathematics, the significance of the fundamental parameters "sideband power content" and "frequency bandwidth." This rather simplified theory shows that well-known practical systems have very different performance characteristics, and we are

prompted to wonder whether there can be an ideal system.

Expanded-band systems of well-known type all show unidirectional improvement of signal-to-noise performance with expansion of bandwidth from that which must be consumed by single-sideband transmission. The thought arises that if the parameter bandwidth can be expanded smoothly and usefully, it would appear inconsistent with nature to set an absolute lower limit to its value for this would probably represent a discontinuity of law. Now that we have established the fact that bandwidth expansion can improve signal-to-noise ratio, is it not conceivable that bandwidth compression can be achieved at the cost of signal-to-noise ratio? In the past, bandwidth compression has been sought partly, of course, to provide a greater speed of signalling in a given frequency band, but also in the belief that the effect of noise would be reduced or eliminated. A fresh examination of the problem, undertaken on the expectation that signal-to-noise ratio will suffer, should have a better chance of success.

The present paper shows that there is no obvious discontinuity in the expansion and reduction of bandwidth for transmission of a given rate of information by telegraphy, and after correlating telegraph and telephone practice, consideration is given to a little-known system, which may become of major importance. In 1938, A. H. Reeves proposed a new method for expanded-band transmission, which he called pulse-step modulation, in which the claim was made that speech may be transmitted entirely without noise at the expense of a constant but very small coding distortion.

This new system of transmission, which has recently received considerable attention by communication engineers and has been variously described as "pulse-code," "pulse-count," and "pulse-digital modulation," has not only a very high theoretical signal-to-noise performance, but also the unique feature of adaptability for both expansion and compression of transmission bandwidth. In the present paper, the suggestion is put forward that step modulation may represent a new and universal ideal system, in which maximum possible benefit can be obtained from the frequency bandwidth available. This conception would not displace the old ideal single-sideband transmission as the latter would remain not only as the particular case in which effective bandwidth available for transmission is equal to the bandwidth of the information wave, but also as the method for achieving a signal-to-noise ratio of unity with a minimum of signal power.

3. Early Principles of Communication

This section will be devoted to a short analysis of simple but important principles of communication that appear to have some significance in building up a co-ordinated theory. These principles are limited in useful application to those systems that were well known before expanded-band systems such as frequency modulation; they are, nevertheless, perfectly sound and need to be well understood before consideration of later developments.

3.1 Keying Systems

3.1.1 *Keyed Carrier Wave*

In the simplest of all methods of transmitting information, the periodic interruption of a carrier wave, noise can always be made to have no effect by the use of a sufficiently small receiver bandwidth and reducing speed of keying until the receiver is able to follow it. The conclusion should not be drawn, as was very common in the early days of telegraphy, that noise can be eliminated by bandwidth reduction, for in this case we leave out the essential parameter of speed of signalling. The correct conclusion, which leaves out no essential parameter, is that for a given factor of safety in output signal-to-noise ratio, the speed of signalling is inherently limited by the amount of noise present in the transmission path.

3.1.2 *Multiple-Position Keying*

If the signal-to-noise voltage ratio of a keyed carrier system is great enough, then simple on-off or two-position keying may be replaced by three-position keying; i.e., on, half-on, and off keying. With sufficient signal-to-noise ratio, this principle may be extended to any desired number of key positions, hence permitting the use of a code that sends a vastly greater amount of information in a given time.

When noise is not a limiting factor, we may use either a code of greater complexity or key at a faster rate and, when noise does not permit a particular code and speed, either speed of transmission may be lowered or a simpler code may be used to provide error-free operation.

3.1.3 *Transmission of Infinitely Variable or Uncoded Waveforms*

The accuracy of transmission of an uncoded waveform can be improved indefinitely by reduction of bandwidth if the speed of sending is correspondingly reduced. Thus, if voice recorded on a gramophone is transmitted as modulation of a carrier wave, noise picked up in the transmission path will have a smaller effect and demodulated signal-to-noise ratio may be improved by reduction of playing speed and receiver bandwidth.

In the case of transmission of a "real" voice at its natural speed, a reduction in bandwidth will result in a corresponding loss in frequency fidelity, so that good signal-to-noise ratio is achieved by direct modulation only when noise conditions are suitable for transmission of a very complex step code of the type mentioned in the previous section.

3.2 Sideband Analysis of Transmission Systems

All signals that are used for the transmission of information can be analysed to yield a system of sidebands, which correspond to the information wave, and other components which carry no information and can be considered as carrier waves. The best-known systems for transmission of speech are, perhaps, the amplitude-modulated carrier wave and the single-sideband system. The former contains a steady carrier wave unaffected by modulation and devoid of any "information" and a pair of equal sidebands that represent the information wave. Single-sideband transmission represents the same transmission with carrier and one sideband completely removed: alternatively, it can be considered simply as the original information wave displaced, positively or negatively, in frequency.

It is convenient to consider some systems as having a "complex" carrier wave, or a whole system of carrier waves. For example, an amplitude-modulated pulse train may be considered as a multiple-carrier amplitude-modulation system. The carrier wave in this case is a constantly modulated wave, which may be analysed as a whole system of "simple" carrier waves: modulation by the information wave produces a pair of sidebands corresponding to each of the simple carrier waves.

For efficiency of transmission, it is evident that the carrier wave or complex carrier wave can serve no useful purpose except the simplification of receiver circuits because, being constant, it could be manufactured and supplied at the receiver. Signal-to-noise ratio of a system using a carrier wave must, therefore, inherently suffer. In some carrier-wave systems, however, the penalty is not great except for very low levels of modulation as the carrier level is a function of modulation level and disappears entirely for the higher percentages of modulation: hence, peak signal-to-noise ratio is not affected.

Transmission systems could be grouped or classified in a number of ways, but the following is found convenient for the purpose of this paper.

A. Systems using sidebands only (e.g., single-sideband and other systems produced as the result of balanced modulation).

B. Systems having a constant carrier wave or a complex carrier wave, plus sidebands (e.g., amplitude-modulated carrier or amplitude-modulated pulse trains).

C. Systems having a single or a complex carrier that varies with modulation depth, plus sidebands (e.g., frequency-modulated carrier, and pulse-time modulation).

As a rule, systems of type *C* have a constant mean power of transmission, the growth of sidebands with modulation compensating exactly for decay of carrier level.

In all these systems, sideband power level rises linearly with modulation power level for very small values of modulation. In systems *A* and *B*, this is true for all values of modulation up to the maximum value, but in systems *C*, non-linearity commences fairly early without producing non-linearity of the overall transmitter-receiver system. Systems of type *C*, if modulated heavily, are outside the scope of the present section, their peculiar characteristics not being appreciated until after practical development of phase modulation.

3.3 Signal-To-Noise Ratio for Systems of Linear Sideband Variation

3.3.1 *Single-Sideband Ideal*

The simplest linear transmission and reception system of perfect efficiency is the single-sideband method. The whole power of the signal is used for transmission of information, and the frequency bandwidth occupied is just the same as that of the original information wave. The mechanism of reception assures equal efficiency of demodulation and, hence, signal-to-noise performance for all signal levels.

Single-sideband performance represents the most satisfactory standard of efficiency for comparison of all other systems, and if conditions of random noise are such that single-sideband working yields a signal-to-noise ratio of less than unity, then *no* system can yield a better performance.

3.3.2 *Ideal Linear Receiver*

The following theoretical method of reception appears to represent an ideal method for all transmission systems of the linear type.

All transmission systems may be analysed to yield a system of sidebands. If at the receiver these sidebands are separately heterodyned, each yielding the original information wave plus noise, the separate demodulated outputs may be added together to yield a combined demodulated output. Such a receiver, which we shall call the "ideal linear receiver", may not be practical, but it is described here as a theoretical mechanism for the study of signal-to-noise ratio.

3.3.3 *Multiple-Sideband Systems*

It may be shown very easily that, applied to a transmission system of multiple sidebands of equal amplitude, the ideal linear receiver yields exactly the same demodulated signal-to-noise ratio that would be yielded by a single-sideband system of equal sideband power. The ideal linear receiver yields the maximum signal-to-noise ratio visualised by the early Carson theory, and still, of course, represents our theoretical ideal for all systems of linear sideband variation.

In a multiple-sideband system, maximum efficiency is not achieved unless all sidebands are of equal level, but if demodulated outputs are suitably weighted before combination, the imperfection factor is very small indeed.

The above conception suggests that high efficiency of a communication system depends upon uniform loading of the frequency band or bands occupied.

3.3.4 *Conclusions*

Single-sideband transmission gives a signal-to-noise performance that represents a theoretical ideal for all systems of linear sideband variation.

Signal-to-noise performance for any system is inherently imperfect according to the ratio of sideband content to total power radiated.

When a signal contains multiple sidebands, the demodulated signal-to-noise ratio does not necessarily suffer by this, but performance can never be improved by radiation of the same mean power as similar sidebands on different frequencies. (It appears from the above that signal-to-noise ratio cannot be improved by the use of expanded-band technique, but it will be shown in a later section that linear systems are, in effect, only equivalent to single-sideband working in true consumption of bandwidth.)

Before continuing to a consideration of non-linear modulation systems, something should, perhaps, be said to clarify the early conception that noise could be reduced by bandwidth reduction. Evidently, noise power picked up in the transmission path may be reduced by reduction of receiver bandwidth, but this is merely a question of adapting the receiver to the required transmission, and reduction below a certain limit must inherently leave out some of the required signal, causing either inability to receive the total information or a deterioration of signal-to-noise ratio. Though the single-sideband method of transmission appears to represent the system that uses the minimum possible bandwidth, it does not necessarily provide a better signal-to-noise ratio than any other system, for a multiple-sideband system of the linear type can give an equal performance.

4. Principles of Expanded-Band Communication

The term "expanded-band" is chosen deliberately to define a transmission system that inherently consumes more effective bandwidth

than single-sideband transmission. Though the term "wideband" has sometimes been used, the latter is considered unsuitable, having been used also in the simple sense of channel-handling capacity.

4.1 General

The demonstration by Armstrong that continuous-wave phase or frequency modulation could, in some circumstances, provide a better signal-to-noise performance than single-sideband technique made a deep impression on the orthodox communication engineer, appearing at first to indicate the necessity for a revision of theory.

The essential difference between systems that are inherently limited to single-sideband performance and those that can give a greater demodulated signal-to-noise ratio has been identified as a difference that corresponds to linear or non-linear variation of sideband power with modulation power level. The reason for superiority of non-linear systems is not, however, easily apparent, and the practical engineer who always expects to have to pay for a radical improvement considers at once what "currency" he must use to "purchase" this radical benefit.

The answer, of course, is that the improved quality of transmission is obtained at the expense of additional bandwidth. Though, at first, additional bandwidth does not always appear to provide improved signal-to-noise ratio, succeeding pages will show that benefit may always be obtained from the *true* consumption of bandwidth.

Confining ourselves for the moment to a consideration of the phase-modulated continuous carrier wave, we see that for small amounts of modulation, bandwidth is constant, but that large modulation produces an increase in bandwidth. For large modulation, bandwidth is proportional to modulation voltage, and so, of course, is the output signal-to-noise ratio.

If we use "bandwidth" as an essential factor for estimation of signal-to-noise performance, we see at once that two simple and practical parameters only—"sideband power content" and "bandwidth"—are required. When modulation is small, signal-to-noise ratio is determined by sideband power content only, and for large modulation, sideband content is substantially constant at 100 percent, when bandwidth variation dominates. This method of analysis is simple and practical, and when applied to any continuous-wave system yields the result that signal-to-noise ratio is proportional to both sideband power content and bandwidth occupied.

4.2 Effective Bandwidth of a Communication System

Effective bandwidth of a communication system is here defined as the true bandwidth divided by the number of times this bandwidth may be used again by other transmissions of the same effective bandwidth without crosstalk between the various systems. Analysis of signal-to-noise performance of communication systems using this new parameter yields the definite result that performance is always dependent upon the effective bandwidth consumed.

In accounting for the performance characteristics of systems using intermittent signals, it is at first tempting to consider effective bandwidth as total bandwidth multiplied by the fraction of time utilised. Such a definition often yields the same result as the one given above but fails to account for the lack of improvement of a number of systems, for example, the amplitude-modulated carrier wave over the single-sideband system.

4.3 Linear Systems Are Not "Effective Wideband" Systems

It can be shown that all systems of linear sideband variation do not effectively consume more bandwidth than the single-sideband method, and as a corollary this means that signal-to-noise performance cannot exceed that of single-sideband working.

It will be noticed that two separate double-sideband systems may be radiated on diphase carrier waves of the same frequency without mutual interference, so that the effective bandwidth of a double-sideband system is not greater than for single-sideband transmission. Taking a more general case, if two sidebands a are radiated on different frequencies in any proportions defined by k_1a and k_2a, then crosstalk does not need to be experienced from another communication channel using the same carrier frequencies, which radiates sidebands b in any different proportions (or phases) defined by c_1b and c_2b.

Separate demodulation of the frequency channels yields (k_1a+c_1b) and (k_2a+c_2b), when a combination in suitable proportions (and phases) can yield outputs containing either a without b, or b without a.

One special case of interest is the amplitude-modulated pulse train. If the information wave contains frequencies between zero and f cycles per second, then the minimum possible pulse-recurrence frequency is $2f$ per second. Now, video-frequency or direct-current pulses which are initiated in a total bandwidth of F cycles per second necessarily have a total duration of $1/F$ seconds, and if it is desired to interlace a number of similar channels with our considered channel, adjacent-channel time spacing without crosstalk may be equal to one half of the pulse width, i.e., $1/2F$ seconds. Such an arrangement would permit the use of F/f channels in the group period $1/2f$ seconds, so that effective bandwidth of a single channel is f cycles, the same as for single-sideband transmission.

Bandwidth consumption of a radio-frequency pulse train is not greater, despite the fact that a radio-frequency pulse of a given duration occupies twice the bandwidth of a video-frequency pulse, because two complete diphase systems can theoretically be radiated on the same carrier frequency.

4.4 Systems of Non-Linear Sideband Variation

A study of non-linear systems shows that effective bandwidth is always greater than for single-sideband transmission, and also that signal-to-noise ratio depends upon the value of this parameter. In many systems, signal-to-noise ratio is directly proportional to effective bandwidth, but the law of improvement is not always the same, as will be shown in the next section.

5. Signal-To-Noise Ratio for Expanded-Band Modulation Systems

5.1 General Method for Derivation of Performance Characteristics

The performance characteristics of various expanded-band systems of transmission may be derived from a study of the particular receiving mechanism used for the system concerned. Obviously, the result obtained must be dependent upon the method used for demodulation, and as it is not always easy to prove that the method of reception must be ideal, it is felt that the method is not suitable for an academic study and comparison of various systems.

Another method has therefore been found, which derives the maximum possible performance figures for any system without reference to any practical receiving equipment. This method has been used to check figures derived by more practical logic, and wherever there is confidence that the practical receiving technique has no fundamental imperfection, agreement is perfect. The following processes are involved.

A. Suppression of modulation to a very small amount, when it may be considered that sideband production is linear with modulation, that effective bandwidth is not greater than for single-sideband transmission, and that sideband power level is less than noise power receivable in that bandwidth.

B. In the above conditions, it is not conceivable that any non-linear technique such as the use of limiters can assist in demodulation of the signal, and we accept Carson's definition of maximum possible signal-to-noise ratio, estimating this from the power of the sidebands present.

C. On increasing modulation level to correspond to practical expanded-band communication, signal-to-noise ratio must, of course, be raised linearly with modulation.

5.2 Signal-To-Noise Performance for Continuous-Wave Phase Modulation

If we phase-modulate a carrier wave by a very small amount, then two sidebands only are produced, and the amplitude of each is $n/2$ times the amplitude of the carrier wave, where n is the deviation in radians.

$$\text{Total sideband power} = 2\times\left(\frac{n}{2}\right)^2$$

$$= \frac{n^2}{2} \text{ times carrier power.}$$

$$\text{Signal-to-noise ratio} = S_{PM}$$

$$= S_{SSB}\times\frac{n}{\sqrt{2}},$$

where S_{SSB} is the signal-to-noise ratio for single-sideband transmission when using the same mean power at the highest level of modulation.

But signal-to-noise ratio is always proportional to deviation, so the formula is correct for all values of deviation.

Thus, $S_{PM} = S_{SSB} \times \frac{n}{\sqrt{2}}$

$= S_{SSB} \times \frac{B}{b} \times \frac{1}{2\sqrt{2}}$ (for large modulation),

where B = bandwidth and
b = bandwidth required for single-sideband working.

The same formula results also from a consideration of the performance of practical receiver circuits, hence confirming that an orthodox receiving technique represents an ideal receiving mechanism.

5.3 Performance of Pulse-Time Modulation

5.3.1 *Effective Bandwidth of a Pulse-Time-Modulation Channel*

We have shown, in Section 4.3, that the bandwidth consumed by an amplitude-modulated pulse train is the same as for single-sideband transmission. In this case, the time allotted per pulse was equal to the pulse rise or half pulse width.

In the present section, we shall consider only the case of pulse-time modulation of large time excursion, so that the time effectively consumed by a single pulse is equal to the total time that may be consumed by displacement modulation of the pulse. We shall neglect the comparatively small time consumed by the unmodulated pulse train.

With this approximation, it is evident that the effective bandwidth of a pulse-time-modulation channel is equal to

$$\frac{\text{total time excursion}}{\text{half pulse width}} \times \text{bandwidth for single sideband,}$$

5.3.2 *Signal-To-Noise Ratio for Pulse-Time Modulation (Video-Frequency Pulses)*

Let time of pulse rise $= t$ seconds, and pulse-recurrence frequency $= N$ per second. Frequency bandwidth $= 1/2t$ cycles per second. Number of harmonics of pulse-recurrence frequency present $= 1/2Nt$. Modulation *total* excursion of t seconds gives a phase excursion of pulse occurrence of $\pm \pi Nt$ radians. Phase excursion of top harmonic

$$= \pm \pi Nt \times \frac{1}{2Nt} = \pm \frac{\pi}{2} \text{ radians.}$$

Now reduce modulation $n\pi/2$ times, where n is very large, when the modulation of the top harmonic $= \pm 1/n$ radian. Each sideband of top harmonic $= 1/2n$ times harmonic amplitude.

Therefore, total sideband energy of top harmonic

$= 2 \times 1/4n^2$
$= 1/2n^2$ times energy of top harmonic.

Now, sideband amplitude is proportional to phase excursion and, therefore, to harmonic order, so sideband energy is proportional to the square of harmonic order.

Sideband-power distribution is therefore parabolic with frequency, and total sideband power equals

$$\frac{1}{3} \times \frac{1}{2n^2} = \frac{1}{6n^2}$$

of total energy of transmission.

At this low level of modulation, where sideband power is linear with modulation power level, we cannot expect a better signal-to-noise ratio than that prescribed as a maximum by Carson.

Hence, signal-to-noise ratio for pulse-time modulation $= S_{SSB} \times (1/6n^2)^{\frac{1}{2}}$ for modulation $2/n\pi$ of pulse rise (where S_{SSB} is the signal-to-noise ratio for single-sideband transmission, using the same mean power at full modulation)

$$= S_{SSB} \times \left(\frac{1}{6n^2}\right)^{\frac{1}{2}} \times \frac{n\pi}{2} \times \frac{\text{total excursion}}{\text{time of pulse rise}}$$

$$= S_{SSB} \times \frac{\pi}{2\sqrt{6}} \times \frac{\text{total time excursion}}{\text{time of pulse rise}}$$

$$= S_{SSB} \times \frac{\pi}{2\sqrt{6}} \times \frac{\beta}{b} \text{ for wideband working,}$$

where β = effective bandwidth and
b = bandwidth required for single-sideband operation.

It may be noted that the constant factor $\pi/2\sqrt{6}$ differs from the corresponding factor of $1/2\sqrt{2}$ for phase modulation and, in fact, represents an improvement of performance of about 5 decibels. Unfortunately, in the practical case of pulse-time modulation applied to radio systems, actual bandwidth occupied by a multichannel system is twice the effective bandwidth,

owing to the fact that the video-frequency system produces two sidebands on the carrier frequency used, so that practical performances of pulse-time modulation and continuous-wave phase modulation are very similar indeed, when using the same mean power of transmission.

5.3.3 *Special Note*

The theoretical method chosen for estimating signal-to-noise performance has found particular value in its application to pulse-time modulation. The equivalent formula developed from consideration of a practical receiver, in which the steep leading edges of pulses are used for demodulation, yields a result which represents an imperfection factor of about 2 decibels. The suggestion is made that such a receiver is inherently imperfect, and that the use of both edges of pulses, on which noise fluctuations are not identical, would assist in bringing performance very near to the theoretical ideal.

Alternatively, frequency-discriminator technique could yield a similar result.

5.4 Performance of Time-Sharing Frequency Modulation

The present section represents a purely theoretical study of signal-frequency-modulated pulse trains, in which instantaneous values of the modulating signal are transmitted as signal-frequency values. Thus, a single communication channel can be considered as a frequency-modulated carrier wave that is "gated" into the form of short pulses. A number of such pulse trains, each bearing a different modulation, may be interlaced to provide a time-sharing multiplex system, and when the various pulses exactly fill up the whole of time, the total transmission is almost identical to that obtained by frequency modulating a continuous carrier wave by a multiplex amplitude-modulated video-frequency train. Thus, the multiplex system is conveniently described as time-sharing frequency modulation.

The apparent super-efficiency of signal-to-noise performance as compared with continuous-wave phase modulation or pulse-time modulation has not, so far as is known to the author, been confirmed by practical development, but there seems to be no reason for doubting it.

5.4.1 *Effective Bandwidth of a Signal-Frequency-Modulated Pulse Train*

In a signal-frequency-modulated pulse system, the pulses are not displaced in time by modulation, and the various pulses of a time-sharing multi-channel system may be permitted to overlap without theoretical crosstalk as were the amplitude-modulation pulses of Section 4.3.

Let the bandwidth occupied by a single radio-frequency pulse $=B_1$ cycles per second.

Let modulation frequency excursion $=\pm B_2$ cycles per second.

Then, total bandwidth $B=B_1+2B_2$.

Pulse rise $=1/B_1$ seconds.

For a modulating or information-wave bandwidth of b cycles, i.e., bandwidth for a single-sideband system, channel pulse intervals $=1/2b$ seconds.

$\therefore$ maximum possible number of channels $=B_1/2b$,

$\therefore$ effective bandwidth β per channel

$$=B\times\frac{2b}{B_1}=\frac{B_1+2B_2}{B_1}\times 2b.$$

Total bandwidth varies upwards with the amount of modulation, and in the special case when bandwidth increase due to modulation ($2B_2$) equals the total bandwidth of the unmodulated system (B_1), effective bandwidth per channel is $4b$, or four times that for single-sideband working.

5.4.2 *Signal-To-Noise Performance for Time-Sharing Frequency Modulation*

Maintaining the annotation of the previous paragraph, the deviation ratio for the highest modulation frequency b is equal to B_2/b.

In this state, all component carrier waves, which constitute the unmodulated repetitive pulse train, are modulated $\pm B_2/b$ radians, and if the amount of modulation is very small, only one pair of sidebands is produced for each carrier.

The power level of each pair of sidebands is $\frac{1}{2}\times\left(\frac{B_2}{b}\right)^2$ times that of the associated carrier power; therefore, total sideband power is $\frac{1}{2}\times\left(\frac{B_2}{b}\right)^2$ times total signal power. For a system

in which deviation ratio is independent of modulation frequency, the signal-to-noise ratio would be

$$S_{SSB} \times \frac{1}{\sqrt{2}} \left(\frac{B_2}{b} \right),$$

but in frequency modulation, deviation ratio is inversely proportional to modulation frequency. This has the effect of providing a triangular signal-voltage pattern or, alternatively, for correct fidelity, a tilted noise characteristic. Noise, therefore, has a square-law power distribution over the modulating-signal band, and total noise power is suppressed three times as compared with the system of constant phase deviation.

Hence, signal-to-noise ratio for time-sharing frequency modulation

$$= S_{TSFM}$$

$$= S_{SSB} \times \frac{\sqrt{3}}{\sqrt{2}} \times \frac{B_2}{b}$$

$$= S_{SSB} \times \frac{\sqrt{3}}{\sqrt{2}} \times \frac{1}{b} \times \frac{B - B_1}{2}$$

$$= S_{SSB} \times \frac{\sqrt{3}}{2\sqrt{2}} \times \frac{1}{b} \times (B - 2nb),$$

where n = number of channels or

$$= S_{SSB} \times \frac{\sqrt{3}}{2\sqrt{2}} \times \left(\frac{B}{b} - 2n \right).$$

Thus, if n is small, performance approximates to that of continuous-wave frequency modulation, despite the fact that only a small part of time is occupied.

In terms of effective bandwidth β,

$$S_{TSFM} = S_{SSB} \times \frac{\sqrt{3}}{2\sqrt{2}} \times \left(\frac{\beta}{b} - 2 \right) \times n.$$

In the special case when modulation is sufficient to double overall bandwidth, when $\beta/b = 4$,

$$S_{TSFM} = S_{SSB} \times \frac{\sqrt{3}}{\sqrt{2}} \times \frac{B}{4b}$$

or

$$S_{SSB} \times \frac{\sqrt{3}}{\sqrt{2}} \times n$$

or

$$S_{SSB} \times \frac{\sqrt{3}}{4\sqrt{2}} \times \frac{\beta}{b} \times n.$$

The last expression indicates an improvement over pulse-time modulation of $(n\sqrt{3})/4$ times, which may not have much significance for a 10-channel system, but would be of vast importance for a 200-channel system.

6. Noise Barrier

The mechanism of reception for any expanded-band system depends for correct operation upon the signal amplitude being greater than the noise amplitude in the transmission medium. For example, in the case of a time-modulated pulse train, peaks of noise equal to signal pulses would cause ambiguity in recognition of pulses, and immediate failure of the receiver. In a similar way, the mechanism for reception of a phase-modulated carrier wave depends upon the carrier amplitude exceeding the noise amplitude at the demodulator.

The effect of such breakdown is to introduce a new type of noise, the amount depending upon the frequency of breakdown.

6.1 Practical Effect of Noise Barrier

In earlier sections of this paper, signal-to-noise performances have been estimated without reference to the breakdown point, which destroys the action of the receiving mechanism, and the impression might be gained that the formulae are quite accurate for moderate expansion of bandwidth, but that they are subject to a simple top-limit of bandwidth defined by peak noise. This, however, is not the case.

First, let us consider the characteristic of noise. For the purpose of the present study, we shall define noise as interference energy uniformly distributed over the frequency band, and which can usually be considered as a spectrum of equal but infinitesimal sine waves of all possible frequencies, spaced uniformly at infinitesimal intervals of frequency.

Such a definition comprises completely random noise, most natural static, and most forms of man-made static, but does not include what might be called "intelligent interference," comprising all interference energy to which the communication system is particularly susceptible, whether it be deliberately designed interference, many types of interfering signals, or some exceptional form of man-made static such as interference from a high-frequency furnace.

Under this definition of noise, though it is true that noise power is always proportional to

bandwidth, the relationship between peak noise voltage (which determines the breakdown point) and root-mean-square noise voltage is not definable. In the case of completely random noise, the maximum noise peak in any bandwidth is theoretically infinite, and in all types of noise it is evident that very high peaks are possible.

In a practical communication system, occasional breakdown cannot be avoided, whatever the bandwidth chosen, but we may deduce the following.

In systems in which output signal-to-noise ratio is satisfactory but in which bandwidth expansion is small, the signal-to-noise ratio in the transmission medium is also satisfactory, so the frequency of breakdown is small, and the new source of noise has little significance. Bandwidth expansion can never reduce the frequency of breakdown, and in most cases will tend to increase it, so that useful expansion of bandwidth is limited. Thus, output signal-to-noise ratio expands with bandwidth, at first according to the law already defined by our formulae, then at a lower rate, usually passing through a maximum value and decreasing rapidly due to very frequent breakdown.

At first sight, consideration of the noise barrier appears to complicate our signal-to-noise calculations to a high degree, but this is hardly necessary in most practical applications of expanded-band systems, for the reason that such systems are normally designed to operate well within the bandwidth corresponding to the noise barrier. It is rare that bandwidth is so plentiful that it can be consumed without reference to the information-carrying ability of the communication channel.

As an example, let us consider the case of a multi-channel pulse-time-modulation system in which 20 separate speech channels are required, the demodulated signal-to-noise ratio of each channel being 60 decibels. The suggestion is made, here, that it would be impracticable to design equipment giving a signal-to-noise improvement over single-sideband working of 50 decibels, for the reason that total consumption of bandwidth would be excessive. It would appear that not more than about 20 decibels improvement over a single-sideband system could be designed for, when the theoretical bandwidth expansion factor would be about 20 for the video-frequency system or 40 for a high-frequency system, and the practical overall consumption of bandwidth is likely to be 80 times the theoretical bandwidth for single-sideband working. A practical bandwidth expansion of 80 times, on a 20-channel speech system of 3000-cycle-per-second fidelity, involves a transmission bandwidth of nearly 5 megacycles per second, which represents a practical receiver bandwidth. In this case, signal-to-noise ratio in the high-frequency transmission channel is about 40 decibels so that the frequency of breakdown due to high peaks of noise is quite insignificant.

In amplification of the above, it may be mentioned that for completely random or fluctuation noise the peak value of noise will exceed the root-mean-square value by a factor of 4.4 for only one part in 100,000 of time.

In cases where total consumption of bandwidth has little significance, the maximum useful expansion depends upon both the type of signal used and the characteristics of the prevailing noise. Significance of type of signal becomes evident when we consider that in a continuous-wave method of transmission the peak signal is independent of bandwidth, but it may be increased very much by the use of intermittent transmission. Thus, frequency modulation is much more susceptible to noise-barrier limitations than pulse-time modulation. The significance of the type of noise becomes evident when we consider that in the case of completely random noise the frequency at which the noise peak exceeds the root-mean-square value by a given factor is statistically definable and is independent of bandwidth. However, in the case of impulse noise, the peak to root-mean-square ratio is proportional to the square root of bandwidth, whereby harmless peaks in a narrow bandwidth become more and more significant with band expansion.

7. *Résumé of Established Principles*

A. Demodulated signal-to-noise ratio or overall quality of transmission corresponding to a given rate of transmission of information, is determined by the signal power available as sidebands and the frequency bandwidth available.

B. In general, increased bandwidth permits improved quality of transmission, but there is a limit to the maximum useful expansion.

C. Expansion of bandwidth towards the limit defined in B introduces a new source of noise, which eventually becomes the predominating factor, and expansion beyond this limit may involve a rapid deterioration of overall performance.

D. Maximum efficiency is obtained by the uniform energy loading of the frequency band occupied.

8. Unification of Telegraph and Telephone Theory

The analysis already given in Section 3.1 shows a very close connection between telegraph and telephone systems. A simple on-off telegraph system is able to transmit only a very simple waveform. The multiple-position keying system of Section 3.1.2 shows that when noise conditions are more favourable, a higher quality of transmission can be achieved by the use of a code of greater information-carrying ability. If the signal-to-noise ratio is large, then the number of possible positions or code units also becomes large, and a more complex sequence of units or a more complex waveform can be transmitted. In transmitting a speech waveform, then, we have in effect an infinite number of code units, and distortion by noise is inevitable.

It was a consideration such as the above which led A. H. Reeves to the conception of transmission of speech without any noise by his step-modulation principle. Before continuing to a description of this system, certain principles will be given that will help in understanding an extension of the simple telegraph to a telegraph system of multiple code units, and then to transmission of *any* waveform such as speech.

8.1 Rate and Quality of Transmission of Information

8.1.1 *Maximum Rate of Signalling*

The maximum number of units of information per second that can be transmitted in a given bandwidth is equal to twice this bandwidth expressed in cycles per second.

This principle is clearly established from the conception of single-sideband transmission, for in the frequency band zero to f cycles per second, any combination of successive amplitudes up to a rate of $2f$ per second may be absolutely defined by a wave that contains frequency components up to f cycles per second maximum. In other words, a Fourier analysis of a waveform that gives an exactly correct result at instants uniformly spaced $1/2f$ seconds does not need to include frequency components higher than f cycles per second.

8.1.2 *Quality of Signalling*

The number of kinds of units of information, for example the number of different sizes of units, that can be transmitted and received without ambiguity depends upon the amount of noise prevailing.

8.1.3 *Production of Noise by Use of Complex Telegraph*

When the number of kinds of units to be transmitted exceeds the value that resolves all ambiguity, a breakdown of the system occurs causing an interference or noise voltage in the receiver, the transmitted waveform being reproduced with errors.

Transmission of smooth and infinitely variable waveforms such as speech involves the effective use of an infinite number of kinds of units of information and thus an inherent production of noise.

8.2 Telephone Transmission Without Noise

If a communication system is capable of producing an output wave in which instantaneous values lie between $\pm V$ volts, when noise voltage peaks cannot exceed $\pm(V/n)$ volts, then an unambiguous code of $(n+1)$ different units can be used. Thus, on transmitting a continuous waveform, this waveform could be first translated at the transmitter to a wave comprised of abrupt steps of amplitude having $(n+1)$ step levels, and circuit noise could never make any of the step levels ambiguous because the differential level of adjacent steps is twice the maximum noise voltage.

In receiving speech transmission by this method, it is not possible for circuit noise to modify the output waveform in any way, so that we can consider reception to be completely

noise-free, at the expense of a small constant distortion in transmission. This represents the basis of the step-modulation principle proposed by A. H. Reeves, showing the final co-ordination between telegraph and telephone systems. This step system will now be considered in greater detail.

9. Step Modulation *

9.1 Suppression of Noise by Quantizing

The first principle of the system proposed by Reeves has already been expressed in the preceding paragraph. The steps must necessarily be imposed on the signal before transmission, rather than at the receiver, to avoid all ambiguity of receiver output due to noise. The receiver, being able to recognise all steps without ambiguity in a pre-arranged series of values in simple arithmetic progression, reproduces the same relative step amplitudes that were transmitted in the form of a series of instantaneous values that are nearest to the correct values of the original infinitely variable wave. Adjacent values are arranged to differ sufficiently to avoid ambiguity in the receiver through displacement by noise: that is, adjacent steps are separated by at least twice the peak noise voltage.

9.2 Coding of Signal Values

To the conception of quantizing of the information wave according to a convenient series of values, a process that may conveniently take place at a uniform sampling rate, Reeves added the principle of more accurate transmission of the waveform concerned by introducing a second channel of equal power, bandwidth, and sampling rate, this channel radiating a new series of step values, which may be used to provide a second-order approximation to the true instantaneous values of the information wave. For example, if the first channel transmits values corresponding to 0, 10, 20, 30, ..., 90, then the second channel transmits a similar series of values that denote discrepancies 0, 1, 2, 3, ..., 9 in the first series. The two channels are thus able to indicate all values between 0 and 99 without ambiguity due to interference by noise. Information-wave distortion may, of course, be reduced indefinitely by the use of additional channels.

At this stage, it became evident that the principle of sending accurate values by means of combinations of values represented a simple coding process; any desired overall accuracy of transmission could be achieved by using a suitable number of channels and step positions on each channel. In considering the maximum extension of bandwidth, as determined by the number of channels, Reeves' original patent makes the suggestions that speech might be transmitted with sufficient accuracy by means of 32 signal values, and that these could be produced by five channels, each of which has only two step positions. In other words, each channel works on a simple-telegraph on-off basis, the total number of combinations being 2^5 or 32.

Speech transmission by means of a binary code of on-off pulses may be achieved in a number of ways. Applications can be envisaged in which the five (or more) on-off channels might be applied to five different wire circuits. Alternatively, the various channels could be allotted to different frequency bands. Most likely, however, the time-sharing principle would be adopted for transmission of not only the code units of a single speech channel, but also the groups of units corresponding to many other speech channels, all in a single wideband channel.

9.3 Step Modulation Is Not a True Coding System

The step-modulation system has been described as a pulse coding system, as combinations of pulses are coded to mean different amplitudes of the signal wave. In this sense, of course, the process of transmission does involve a coding operation, but the coding operation is very different from the type of coding that has previously been proposed for the transmission of speech.

Just as a specialised letter code is only applicable to a particular language and even to a specialised field of interest, coding systems for speech, which have already been given that designation, have been dependent upon certain assumptions about the waveforms that could possibly be produced by the human vocal cords. The point to be made here is that the

* The present paper was written in original form in October, 1946, when the author had no knowledge of any practical development of pulse-code modulation, as the system is known today.

step system is capable of transmitting any kind of waveform, including all kinds of sound waveforms whether harmonious or otherwise.

Though the steps imposed upon the signal provide a semblance of coding of the type with which we have previously been familiar, the distortion produced may be controlled to any degree of precision required by the information.

In actual practice there is no necessity to tolerate the presence of any coding distortion. The author has suggested that if residual "true" noise is preferred to circuit distortion, this is achieved simply by radiating the channel that transmits the smallest signal steps, not in the form of a finite series of specific values, but in the form of the exact discrepancy of signal amplitude defined by the other channels. In this case, there can be no trace of a step in the demodulated signal, the circuit noise of the last or maximum-detail channel contributing the exact amount of signal required, plus a little noise just as in other systems of transmission.

10. Signal-To-Noise Performance of Step Modulation

Signals transmitted by step modulation can be considered as having no noise, but a constant circuit distortion, or as having a distortion noise level, the peak of which cannot be greater than half the difference between two adjacent step levels. As a rule, the type of distortion produced can be more easily tolerated than random noise of the same power level, so that if we treat the step distortion as noise, we shall appear to be making a conservative estimate of signal-to-noise performance. However, full advantage cannot be taken of the smaller nuisance value of step distortion, for if we use this to reduce the total number of signal values, during periods of low modulation level, percentage distortion is liable to rise by an intolerable amount. To avoid this defect, practical development has introduced the feature of logarithmic amplitude compression of the information wave before quantizing so that successive steps correspond to information-wave values arranged in geometric progression. For example, if a total of 32 signal values are used, 16 positive and 16 negative, these may be chosen at 1.5-decibel intervals, but converted to linear steps of approximately 6 per cent of maximum amplitude by amplitude compression. In such a system, the amount of distortion or step noise varies with signal output, but percentage distortion is almost independent of modulation level.

The present study will be confined to a consideration of coding of signal values arranged in simple arithmetic progression. In this case, though step noise is zero for the condition of no modulation, it is otherwise almost independent of modulation level, and comparison with other expanded-band systems is made easier.

First, however, to make a comparison of signal-to-noise performance upon an identical basis, we shall consider performance of the modified step system mentioned in the previous section. In this system, in which the maximum-detail channel is transmitted as an exact correction, noise has the same meaning as for normal transmission systems, and comparison is completely fair.

10.1 Signal-To-Noise Performance of the Modified Step System

In the present analysis, it will be assumed that the various coding channels are each transmitted by single-sideband methods.

On dividing the available signal power into n equal parts for division among n separate channels, maximum channel signal-to-noise ratio $= S_{SSB}/\sqrt{n}$.

This value represents the signal-to-noise ratio of the fine-detail channel. However, we have $(n-1)$ other channels, which may each comprise signal values corresponding to any of N information-wave values, the number of step positions chosen for each of these channels.

These $(n-1)$ channels have a total of N^{n-1} code positions, and these are effective in expanding the signal-to-noise ratio of the fine-detail channel by the same amount.

Hence, signal-to-noise ratio for the step system

$$= \frac{S_{SSB} \times N^{n-1}}{\sqrt{n}}.$$

In this formula, n is the bandwidth expansion factor β/b used in analysis of other systems. The important feature to be noticed is that demodulated signal-to-noise ratio is not a simple function of bandwidth expansion as in the case of frequency and pulse-time modulations, but

can be made very large if N can be large. The value of N, the number of step values, is of course limited by the noise-barrier conditions, and the exact limitation must depend upon the nature of prevailing noise.

10.2 Consideration of Noise Barrier

The step system appears to be different from any other transmission system in that constants can be chosen that make the noise barrier or breakdown threshold a very serious consideration, even when only a very moderate bandwidth-expansion factor is used. In other words, the enormously improved suppression of the only type of noise that can have much practical significance in most transmission systems is attained only at the expense of reduced factor of safety against what has previously been considered as a breakdown effect. However, it is usually found that a design that has equal limitations imposed by two or more separate effects is more economical than one which accidentally avoids all but one of such limitations.

It is beyond the scope of the present study, to consider the exact optimum design for a step-modulation communication system, in which normal and breakdown noise impose equal limitations, but approximations will enable us to draw useful conclusions.

As previously stated, in the case of completely random noise, the peak value will exceed the root-mean-square value by a factor of 4.4 for only one part in 100,000 of time. For communication in which ultimate signal-to-noise ratio is moderate, the occasional breakdown caused by peak noise can be permitted, and we may assume an *effective* peak to root-mean-square ratio for noise that is not greatly in excess of 4.4.

In considering the performance of step modulation, this factor becomes significant in a determination of the maximum number of unambiguous values that can be transmitted in a single channel, and also in the number of coding channels that can be used, causing the noise barrier to assume a dominating importance despite the use of a small frequency band. We, therefore, introduce the factor p to define the ratio of effective peak to root-mean-square noise in a transmission bandwidth equal to that of the information wave.

10.3 Variation of Signal-To-Noise Ratio with Bandwidth-Expansion Factor

On splitting the available signal power into n equal parts for division among n separate channels, channel maximum signal to root-mean-square noise ratio $= S_{SSB}/\sqrt{n}$ and maximum signal-to-*peak*-noise ratio $= S_{SSB}/(p\sqrt{n})$.

In each channel, the available power may be used to provide N instantaneous values, this being achieved without ambiguity provided that

$$N < \left(\frac{S_{SSB}}{p\sqrt{n}} + 1\right).$$

The total number of combinations of the simultaneous n values $= N^n$, and the demodulated maximum signal-to-peak-noise ratio $S = N^n - 1$.

The maximum value of this function corresponds to the maximum value of N, which is the integral value between

$$\frac{S_{SSB}}{p\sqrt{n}} \quad \text{and} \quad \left(\frac{S_{SSB}}{p\sqrt{n}} + 1\right).$$

Therefore, maximum possible signal-to-*peak*-noise ratio S approaches very near to

$$\left(\frac{S_{SSB}}{p\sqrt{n}} + 1\right)^n - 1.$$

This function is plotted in Fig. 1 for various values of S_{SSB}/p.

The expression is valid for both the normal step-modulation system, and also for the modified system in which the fine-detail channel is transmitted as an exact value. In the latter case, however, effective peak to root-mean-square noise is equal to p, so we can state that maximum possible signal to root-mean-square noise ratio is not greater than

$$p \times \left[\left(\frac{S_{SSB}}{p\sqrt{n}} + 1\right)^n - 1\right].$$

In this expression, the number of channels n represents the bandwidth-expansion factor β/b, as used in earlier sections.

10.4 Maximum Signal-To-Noise Ratio for a Given Available Power

The expression derived in the last paragraph appears, at first sight, to permit unlimited expansion of bandwidth and signal-to-noise ratio.

This, of course, is not true, the formula being valid only when channel signal-to-noise ratio is greater than unity. Thus $S_{SSB}/p\sqrt{n}>1$.

Hence,

$$n_{max}=\left(\frac{S_{SSB}}{p}\right)^2$$

and maximum signal-to-peak-noise ratio

$$S_{max}=2^{(S_{SSB}/p)^2}-1$$

$$=2^{n_{max}}-1.$$

The function S_{max} is plotted in Fig. 1 against n_{max} and this serves to show the maximum expansion of bandwidth possible, and the maximum possible signal-to-noise ratio obtainable for a given available power of transmission.

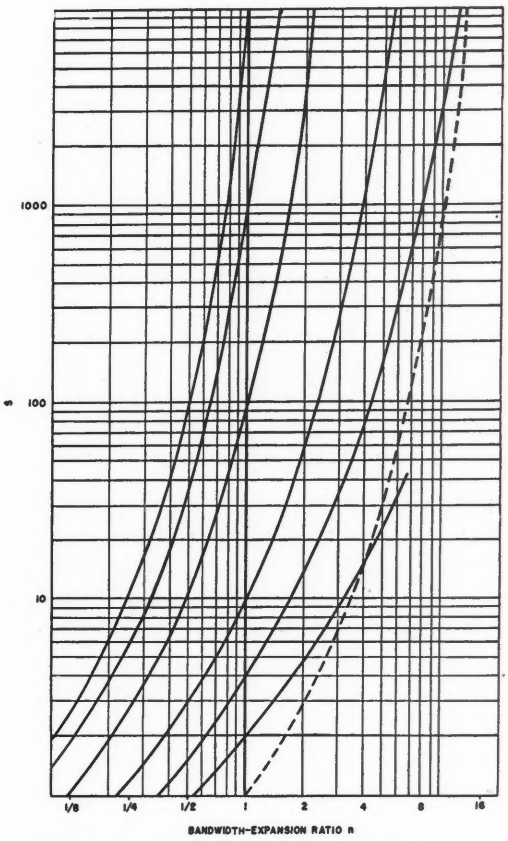


Fig. 1—Performance curves for step modulation.

$$S=\left(\frac{S_{SSB}}{p\sqrt{n}}+1\right)^n-1.$$

$$S_{max}=2^n-1.$$

When $n=1$, $S=S_{SSB}/p$ as for single-sideband transmission.

11. Bandwidth Compression

11.1 Theoretical Possibility of Bandwidth Compression

The possibility of coding speech so that the normal information or meaning may be transmitted in a smaller bandwidth is very well established. Such methods, however, either leave out some characteristics of the voice, which may not be of importance in the sense that exact accent may not be required, or else the speech waveform may be assumed to have particular characteristics that are specially favoured by the transmission method chosen. This technique is outside the scope of the present study and cannot be considered, academically, as better than bandwidth reduction by reduction of frequency fidelity.

Early attempts at true bandwidth compression have usually been undertaken with the object of noise reduction, and this may be why no success has been achieved.

Earlier sections of this paper have established that signal-to-noise ratio may rise rapidly with increase in bandwidth, and also that with a given amount of power and bandwidth, signal-to-noise ratio can be raised by reduction of rate of transmission of information. Surely, then, if we increase rate of transmission of information in a given bandwidth, or if we maintain the rate of transmission in a reduced bandwidth, we must expect to reduce signal-to-noise ratio.

It is evident from earlier sections of this paper that rate of transmission of information is not inherently limited by bandwidth. The amount of information that can be transmitted, when using a suitable code, depends not only upon the number of units of information per second (defined by bandwidth), but also upon the number of kinds of units, defined by signal-to-noise ratio. To transmit a waveform such as speech through a narrow band, we must evidently modify our speech waveform to have a smaller number of units per second, and to compensate for this, the quality or number of kinds of units must be increased.

This is the statement of principle expounded by the author in a discussion with A. H. Reeves in 1941, when the latter at once suggested, "Surely, then, all we need is the exact inverse of the step system," and proceeded to explain the exact mechanism for inversion.

Stated simply, if signal-to-noise ratio is such that by single-sideband transmission, 100 separate values can be transmitted without ambiguity, then these values could be coded to indicate all possible combinations of pairs of values of a set of 10 possible values. Thus, the rate of transmission of units may be halved if quality of unit is allowed to suffer. By this mechanism, bandwidth may be halved if signal-to-distortion ratio is allowed to fall from approximately 40 to about 20 decibels.

11.2 Inverted Step System

An examination of Fig. 1 shows signal-to-noise curves extending on both sides of the single-sideband abscissa of bandwidth. This extension will be found to be perfectly valid if the horizontal scale is considered to be marked in effective bandwidth. In other words, if actual bandwidth is maintained constant and only half the pulse periodicity is used, effectively filling only half of time, then no assumptions are necessary for establishment of validity.

If it is desired actually to transmit a waveform in truly reduced overall bandwidth, we must pause to consider the effect of reduction of bandwidth on effective noise peak voltage.

For purely amplitude-modulated noise or impulse noise, the noise peak is proportional to bandwidth, and such an assumption would permit a better signal-to-noise ratio than is shown by the curves for bandwidth compression. For purely frequency-modulated noise, by which we mean noise concentrated in a very small bandwidth and very slowly sweeping over the frequency spectrum, then noise peak is substantially independent of bandwidth, and reduction of signal-to-noise ratio must be more serious than the curves indicate. For completely random noise, in which effective noise peak is proportional to the square root of bandwidth, the curves are valid for expression of signal-to-noise variation with true physical bandwidth.

The above conception does not, of course, deny the possibility of true bandwidth compression without assumptions with regard to noise, but a new set of curves would have to be produced, showing a rather more pessimistic performance for bandwidth compression. For the purpose of the present paper, we confine attention to effective-bandwidth reduction in which true bandwidth is not reduced, but the number of channels in a given bandwidth may be increased without theoretical limit.

The theoretical possibility of bandwidth compression is interesting from an academic standpoint, but it is very doubtful whether it can provide a useful engineering result. Quite apart from the difficulties involved in production of practical equipment, current problems would appear to favour a slight increase in bandwidth from the basic single-sideband system, rather than reduction, to provide satisfactory signal-to-noise ratio with economy of power.

12. Step Modulation as an Ideal System

Examination of signal-to-noise formulae developed for various communication systems suggests that there are two systems of outstanding efficiency of transmission, in their ability to use a given power and effective bandwidth to set up the greatest signal-to-noise ratio at the receiver output. These systems are the intermittent frequency-modulation system of Section 5.4 and the step-modulation system.

In terms of effective bandwidth, the intermittent frequency-modulation system appears to give the better result in some circumstances, but in terms of the same total bandwidth the step system is always superior. The frequency-modulation system appears superior (for constant effective bandwidth), if the prevailing noise is of the type in which effective peak to root-mean-square ratio is not increased by bandwidth expansion, as is the case for completely random noise. In the case of impulse noise, however, in which noise peak is proportional to bandwidth, the frequency-modulation system is very limited in expansion of total bandwidth.

Confining attention to the use of the same total bandwidth, it seems certain that the step system represents the most efficient system known, and in view of the fact that it appears to utilise every possible feature of communication efficiency, one is tempted to wonder whether it

may represent an ideal that could never be surpassed. In support of this suggestion, one may list the following features.

A. The whole signal power is used to produce sidebands or information-bearing power.

B. Effective bandwidth is controllable for adaption to noise conditions.

C. Signal power uniformly loads the frequency band occupied.

D. Signal power uniformly loads the dimension time, hence minimising total bandwidth.

E. Given a fixed bandwidth, constants may be chosen so that normal noise and breakdown noise impose equal limitations.

F. It is interesting to consider the feature of reversibility of the process of bandwidth expansion and compression. It is likely that the most efficient system possible is capable of inversion, yet if so, there is only one reversible system. Inversion, without grave discontinuity, of a system that gives an inferior performance on one side of the single-sideband standard would give a better performance than the ideal on the other side of this abscissa. Thus, if the ideal system is reversible, there is only one reversible system, the step system.

In conclusion, though considerable evidence tends to set up the step-modulation system as a system of very high theoretical performance, it must not be assumed that it may eventually take the place of all others. Even if we confine attention to theoretical merit, the system does not dislodge single-sideband transmission from its position as a theoretical ideal, except when bandwidth available for transmission is in excess of single-sideband requirements. The single-sideband system can be considered as a special case of step modulation, in which the bandwidth consumed happens to equal the information-wave bandwidth. It may be that nature's choice of frequency bandwidth for the human voice is very fortunate, and that any highly developed communication system will find that this represents a natural optimum.

In considering practical difficulties, we may note that the single-sideband system was unable to dislodge the use of simple amplitude modulation from many applications, and that the considerable complexity of signal-coding devices may well retard or prevent the use of step modulation even in cases where large bandwidth is available. We can only say, with certainty, that in any case where available bandwidth is very large, theoretical performance of step modulation is vastly superior to that of single-sideband working.

13. Acknowledgment

The author wishes to acknowledge valuable criticism by Messrs. B. B. Jacobsen, K. G. Hodgson, V. J. Terry, and others, together with suggestions for expression of thought in a manner likely to achieve a wider range of appeal.

New Developments in Marine Radio Direction Finders*

By H. BUSIGNIES

Federal Telecommunication Laboratories, Incorporated, Nutley, New Jersey

A REVIEW of the past and present state of marine direction finding and a comparison of its merit and fields of application with those of radar and loran are presented. Although marine direction finders have incorporated all of the design and component progress made in radio in the last 20 years, they are still not only working on the same basic principles, but the means of achieving measurements are practically identical to those used 20 years ago. Some of the reasons for this condition and factors that seem to limit further progress are reviewed.

A number of suggestions for improvement are: use of other parts of the frequency spectrum, use of pulse transmissions to reduce night effects, direct-reading indication of bearings, and combination of distance measurement with bearing measurement to obtain a fix from only one beacon station.

• • •

It seems important to compare the marine direction finder with the new radio aids to navigation that were developed during the war, as a means of determining the causes for what seems to be an over-stabilization of design; and, as indicated by the title of this paper, to survey possible new developments in the field that may make the marine direction finder a still more useful instrument of navigation.

The technique of the simple loop direction finder operating on medium frequencies is now too well known to justify repetition in this paper. Also well known are propagation effects,[1] deviations caused by the ship's structure, and deviations due to other causes.

1. Evolution and Present State of the Art

The direction finder is the oldest radio aid to navigation. In the form of either the small rotatable loop or, before that, the large, fixed, crossed loops, the direction finder was the first instrument making use of the directive properties of electromagnetic waves to determine the direction of arrival or origin of a radio transmission. Early disclosures before World War I give the names of Stone, Bellini-Tosi, and Blondel as among the first to invent, propose, and try radio direction-finding systems. Later, Watson-Watt, Mesny, Adcock, Smith-Rose, and many others continued to establish the foundations of the new science by studying propagation and polarization effects, deviations, reflections, and other transmission vagaries.

Both fixed and mobile direction finders have been or can be designed to cover all of the frequency spectrum in which radio transmissions are made. Indications are very often obtained in a direct-reading manner, as through a mechanical pointer on a fixed scale (aircraft radio compass),[2] or in an instantaneous manner on the screen of a cathode-ray tube (fixed high-frequency direction finder).[3,4]

The marine direction finder for shipboard installation became practical with the availability of radio amplifiers after World War I. In the United States, the early work of F. A. Kolster, who in 1913 was already advocating the installation of shore radio beacons, resulted in a demonstration of the radio compass and position finder on board the U. S. Lighthouse Tender *Tulip* in

* Presented before the International Merchant Marine Radio Aids to Navigation Conference, New London, Connecticut; April 30, 1947.

[1] H. Busignies, "Evaluation of Night Errors in Aircraft Direction Finding, 150–1500 Kilocycles," *Electrical Communication*, v. 23, pp. 42–62; March, 1946.

[2] H. Busignies, "The Automatic Radio Compass and Its Applications to Aerial Navigation," *Electrical Communication*, v. 15, pp. 157–172; October, 1936.

[3] "Electrical Communication: 1940–1945," *Electrical Communication*, v. 23, pp. 214–216; June, 1946.

[4] H. Busignies, "Applications of High-Frequency Solid-Dielectric Flexible Lines to Radio Equipment," *Electrical Communication*, v. 22, n. 4, pp. 295–301; 1945.

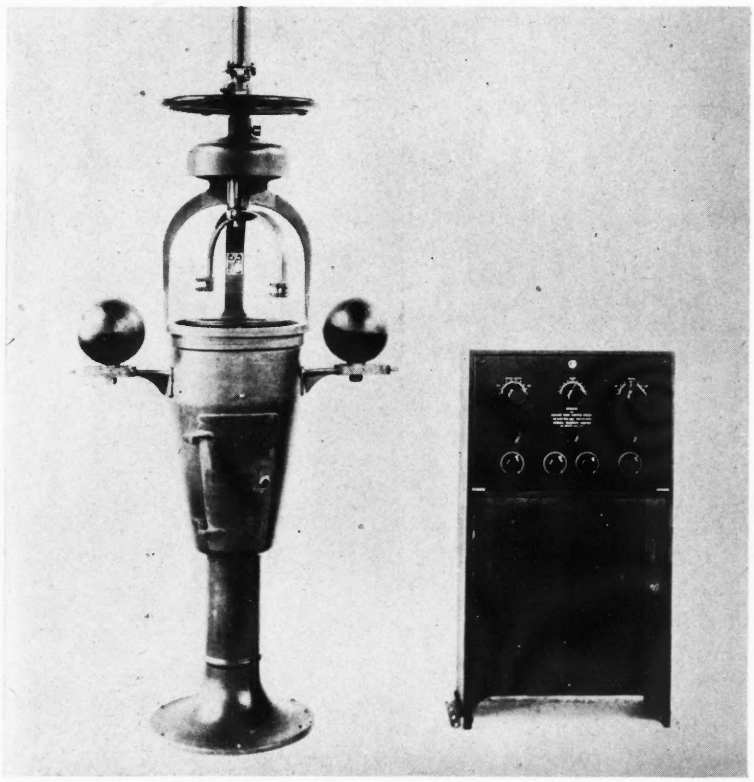
June, 1921.[5] In 1923, marine direction finders were placed on the market by the Federal Telegraph Company of Palo Alto, California.[6] They consisted, basically, of a rotatable shaft that carried a loop and a pointer moving in front of a magnetic-compass dial. Remarkably enough, despite all progress made in radio since that time, all marine direction finders now on the market in the United States consist basically of hand-operated rotatable loops; the pointer actuated by the loop shaft moves in front of a fixed scale or a scale controlled by a repeater from the gyro compass.[7] These direction finders cover the frequency range of 285–315 kilocycles per second assigned to this service. The frequency range is often extended to 500 kilocycles to permit taking bearings on ships and fixed stations transmitting within this extra band. The bearing of the radio beacon is determined from the signal nulls corresponding to the well-known figure-of-8 directivity pattern of the loop collector. Proper design, including balance of the loop and sensitivity of the receiver, determine the sharpness of the nulls and the quality of the direction finder. Deviations caused by the ship's structures, which may run as high as ±15 degrees, are compensated by a mechanical corrector in the form of a cam adjusted during calibration. The elimination of sense ambiguity is obtained through the transformation of the figure-of-8 diagram into an approximate cardioid diagram by combining the loop signal with that from a small vertical antenna.

The modern marine direction finder of the rotatable-loop type, using either a tuned-radio-frequency or a superheterodyne receiver, is a reliable instrument which, when properly calibrated and maintained, will provide readings with an accuracy of ±1 degree over 70 or 80 percent of the bearings, and 2 or 3 degrees for the rest of the bearings. Ranges up to 100 or 150 miles on medium-power beacons are currently obtained. Recalibration after months of use will not show more than 0.5- to 1-degree

[5] F. A. Kolster and F. W. Dunmore, "The Radio Direction Finder and Its Application to Navigation," *Scientific Papers of the Bureau of Standards*, No. 428; January 16, 1922.

[6] F. J. Mann, "Federal Telephone and Radio Corporation," *Electrical Communication*, v. 24, pp. 396–397; December, 1946.

[7] E. H. Price and W. J. Gillule, "Marine Navigation Aids," *Electrical Communication*, v. 22, n. 1, pp. 56–69; 1944.



First marine direction finder placed on the market by Federal Telegraph Company for commercial use in 1923. A similar instrument was demonstrated by F. A. Kolster in June, 1921, on board the U. S. Lighthouse Tender *Tulip*.

difference from previous calibrations, if the superstructure of the ship near the loop is not modified.

Abroad, the principle of fixed loops and rotatable goniometers of the Bellini-Tosi type has often been used, as well as the rotatable-loop principle.

Efforts were made as early as 1926 to produce automatic direction finders requiring only tuning of the receiver to the desired radio beacon for the position of a needle on a scale to indicate the bearing directly. In 1929, a direction finder of the automatic type, developed by the laboratories of the International Telephone and Telegraph System in Paris, was demonstrated on a small private ship in Mediterranean waters. It furnished a direct 360-degree indication without sense ambiguity. A few years later, about 25 such automatic direction finders were installed for normal service on ships of the Portuguese navy. While considerable maintenance was required, the principle of automatic operation without sense ambiguity was completely achieved. Despite these early efforts, no marine direction finder of the automatic type is presently in practical use, so far as current information indicates. This is undoubtedly due to two causes:

A. The apparatus built by the first experimenters was costly and probably too complicated for normal maintenance.

B. The simplicity of operation and reliability of the rotatable-loop direction finder was such that there was very little incentive to change to the automatic type.

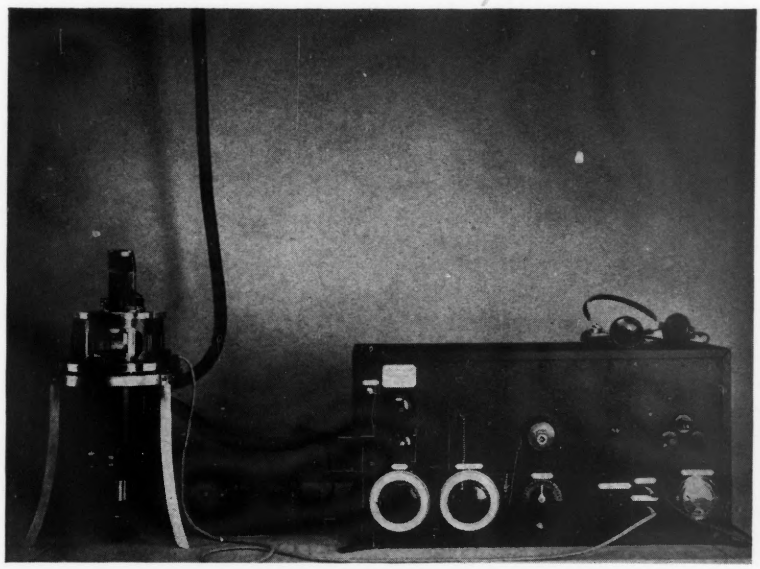
Simplicity is probably also the factor that limited the application of more complex radio beacons, such as those using pulse transmissions, to provide discrimination between direct and ground waves, thus eliminating part or all of the night error that limits the direction-finding range. Considerable research and design work in high- and ultra-high-frequency direction finding was done prior to and during World War II. This resulted in a considerable improvement in the operation of fixed high-frequency direction finders. During the war, the United States Navy installed an important net of such stations, which are now operated by the U. S. Coast Guard.

The Army Air Forces and the Air Transport Command also installed hundreds of instantaneous cathode-ray-indicator direction finders, primarily for guiding and rescuing aircraft, before any other long-range navigational system was available. On the eastern seaboard of the United States, a net of SCR-291 direction finders can be alerted in case of need. Medium-frequency Adock, fixed, direction finders capable of high accuracy at night are also available.

Present type direction finder for merchant marine use.

The night effect is the main limitation on medium-frequency direction finders of the loop type. Direction finders, of course, are not the only radio aids to navigation that suffer from the

effects of sky waves reflected by the *E* and *F* layers of the ionosphere. Solutions to the problem of night effect that have been applied to shore direction-finding stations have not, so far, been applicable to shipboard installations, because of the size and disposition of the antenna system. Night range is, therefore, limited to something of the order of 50 miles. In the last part of this paper we will return to this problem and discuss possible improvements.



First automatic direction finder developed for experimental work in 1928 by the Paris laboratories of the International Telephone and Telegraph System, and tried out on a ship the same year. This direction finder indicated spontaneously on a moving scale the direction of a radio transmission without any ambiguity and throughout 360 degrees. It was the first automatic direction finder developed.

In summary, the first radio aid to navigation, the marine direction finder, is a reliable and simple instrument, still making use of the rotating loop and a means of indication of the same type as was used 20 years ago.

2. *Comparison of Merit and Field of Application with Radar and Loran*

As a result of the inventions and progress made in such pulse-timing systems as radar and loran during World War II, the marine direction finder seems to be subjected to considerable competition. The opinion that the direction finder might become unimportant gained some momentum at the end of 1945 and during 1946, when the first information on the new radio aids was released. At the present time, the situation is much clarified; the performances of radar and loran are well understood. A detailed examination of fields of application shows that marine direction finders, marine radar, and loran are complementary systems, solving different problems of navigation with advantageous overlapping in some cases.

Marine radar is a short-range radio aid to navigation, primarily useful for anticollision service, and for pilotage near the shore or at the entrance to and inside of harbors. A range of 10 to 20 miles (16 to 32 kilometers) is a reasonable figure for its use in these applications.

Loran, on the other hand, is a medium-range (400-mile or 645-kilometer) accurate navigational system, the shipboard operation of which requires somewhat more time and more training than the direction finder. The main question is one of world availability and standardization of loran signals, as operators will have to contend with stations installed in foreign territories in many instances.

Therefore, the marine direction finder, with a day range of the order of 150 miles, is a very excellent intermediate-distance aid to navigation that can bring a ship, with desired accuracy, within the range of operation of marine radar. The advantage of the direction finder for intermediate-distance navigation is simplicity, ease of maintenance, availability, and the relatively small cost of the radio-beacon ground stations. About 190 radio beacons are installed on the shores of the United States and around the Great Lakes. Another 150 are operating throughout the rest of the world. Many countries

can increase the services made available at their shorelines at moderate expense.

We shall conclude this section by repeating the more or less accepted opinion that marine direction finders, radar, and loran are complementary systems; that availability of loran transmissions will certainly be the largest factor in determining its generalized use; and that the operation of radar in collecting its information independently of ground or other ship cooperation is the essence of its assured success.

3. Present Crystallization of Design

In the first section, showing briefly the evolution and the present state of the art, it was stated that the present marine-direction-finder designs are more or less crystallized around the principle of the hand-controlled rotatable loop, and that the means of measurement by observation of the nulls of the figure-of-8 directive pattern is the same as was used 20 years ago, despite the extraordinary development of radio communication and radio navigation during the same period. The art of direction finding as a whole has progressed during the war to the point where techniques and instruments are available or could be made available in the entire frequency spectrum of radio waves; studies have also been made that greatly improved our knowledge of propagation. Indicating means have been considerably improved; the automatic aircraft compass equips almost every aircraft, the large airplanes even carrying two or three such instruments. Shore direction finders employ cathode-ray tubes for instantaneous presentation of bearings. Ultra-high-frequency direction finders are used in connection with airport traffic control. Moreover, investigation of the applicable patent literature is most revealing; the field of direction finders has not been neglected by inventors. There seems to be no relation between the crystallization of design and the very large number of systems that have been invented. We shall list below the factors that seem to limit introduction of new designs and ideas in marine direction finding:

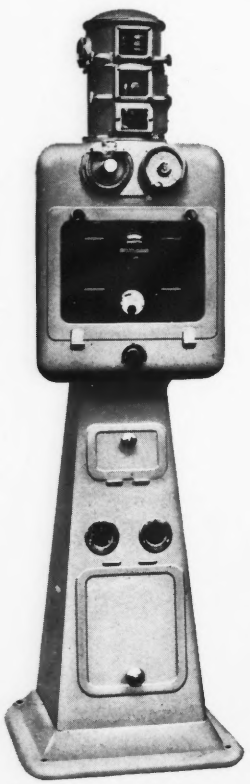
A. Reliability and simplicity of the present design.

B. Availability of a world-wide system of marine radio beacons.

C. Simplicity of the ground beacons and their consequent inexpensiveness.

D. Reluctance and difficulty of introducing changes in the existing world-wide radio-beacon setup.

F. Also, perhaps, the concentration of the attention of laboratories and manufacturers on the problems of radio aids to air navigation.



A commercial version of the completely automatic marine direction finder installed on a number of ships between 1931 and 1934, developed and built by the Paris laboratories of the International Telephone and Telegraph System.

4. New Approaches by a Definite Departure from Existing Solutions

Let us next consider whether we have reached a situation where further progress can be obtained at reasonable cost, or whether some new approaches can make the direction finder a still more useful and attractive instrument for navigation.

4.1 Frequency

It might be questioned whether the present frequency range of 290–310 kilocycles is the most advantageous for this service, or could the use of another part of the frequency spectrum result in improved performance with respect to accuracy, range, or reduction of the night effect?

Lower frequencies are eliminated because their inherent characteristics increase the cost of ground stations and the pressing demand for such frequencies by other services; therefore, higher frequencies are suggested. Experience in the field of shipboard high-frequency direction finders leads us to reject that portion of the spectrum for this application. Ultra-high frequencies would definitely be free from night

effect, but the range would be limited to line-of-sight conditions; the ultra-high-frequency wave collector would have to be installed on the top of the ship's mast to obtain a 20-mile (32-kilometer) range from a transmitter a little higher than sea level. Despite the increase in accuracy and the elimination of night effect, it seems quite improbable that these would compensate for the increase of cost and, more especially, for the shortening of the range. The incentive for any group to develop an ultra-high-frequency direction-finding system on such a fragile foundation is certainly too weak for a manufacturer to experiment in that direction.

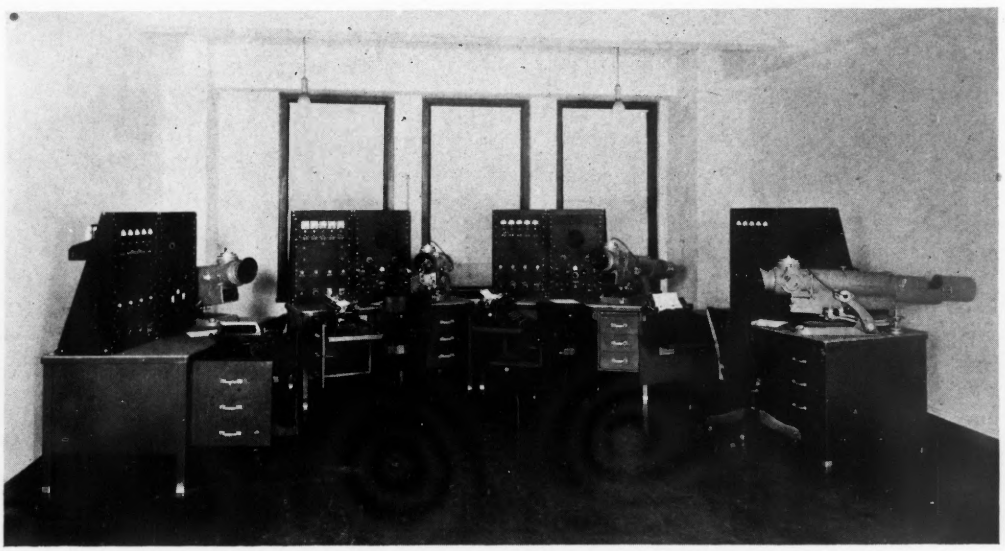
4.2 Instantaneous Direct-Reading Indicator

Would an instantaneous direct-reading indicator of simple design sufficiently increase efficiency to make it attractive to ship operators and owners, and subsequently justify the effort necessary for its development and adoption?

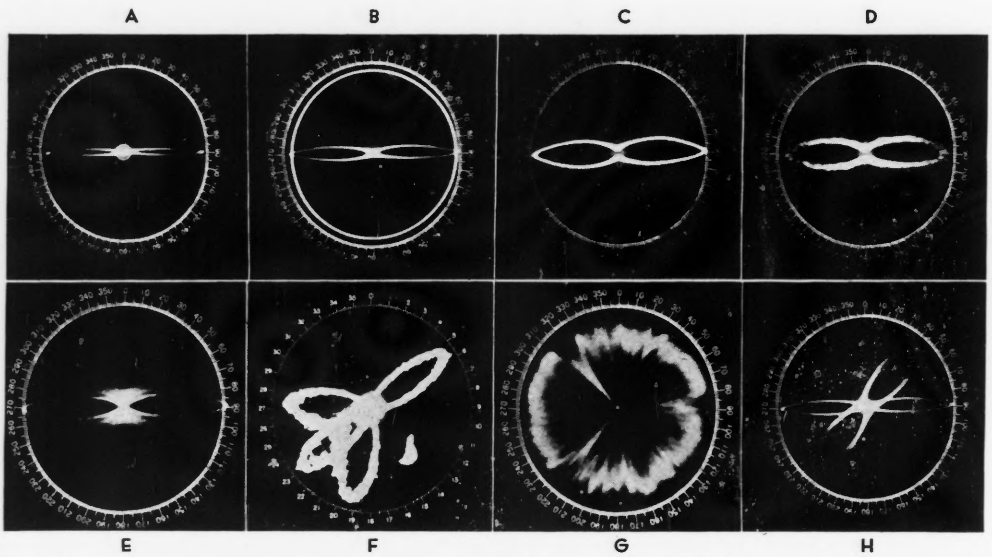
Navigation along the eastern coast of the United States, where beacons are arranged by groups of three on the same frequency and transmit sequentially by units of three minutes, tends to demonstrate that an instantaneous indicator, giving successively the bearings of the three stations, would be quite justifiable on the basis of simplification of procedure. A computer could be designed to facilitate either the recording or the plotting of the fix. The type of indication should be such that a failure would be immediately detectable. It could be a mechanical pointer, a neon or other gas tube, or a cathode-ray tube; however, the simplest and most reliable instrument is to be preferred. In connection with such an indicating system, a null-seeking loop or goniometer, or a small continuously rotating loop or goniometer, is generally required.

While a complete direction finder of such a type has not yet been installed for service tests on board a merchant ship, various components performing equivalent functions have been tested during other developments, and results proved their reliability. This will be an important factor in the choosing among the many possible solutions that offer themselves for direct-reading direction finding. The preferred solutions will be the use of instruments for which a good quantity of data on reliability is already available. Among such we find the null-seeking type of direction finder with mechanical pointer, the rotating loop or goniometer with cathode-ray-tube presentation, and the neon-tube or meter indicators.

Other methods (those avoiding rotating parts) involve electronic goniometers or various



Fixed shore station of high-frequency direction finders for aircraft operated by the U. S. Coast Guard. This direction finder makes use of instantaneous bearing indicators of the cathode-ray type.



Typical patterns of bearing indication shown under various conditions by the cathode-ray direction-finder indicator. *A*—Strong continuous-wave signal. *B*—Strong keyed continuous-wave signal. *C*—Weaker signal. *D*—Weak signal with noise. *E*—Modulated signal. *F*—Superimposed "direction" and "sense" patterns for weak signal with noise. *G*—Noise only. *H*—Two simultaneously keyed signals.

modulation schemes that eliminate the need for three-channel receiving systems.

Generally speaking, whatever the method used to obtain the directive information through one receiving system, the indicator can be of any of the three most important classes just mentioned.

Thus, there is room for ingenuity in design to satisfy the requirement for simplification of procedure.

4.3 Frequency Stability and Bandwidth

Could more frequency stability in the transmitter and receiver permit utilizing channeling principles, as in communication systems, or as proposed for other navigational systems?

The resulting simplification of the tuning of the receiver to the radio beacon should rightly be an added feature of a direct-reading direction finder. Stability has already been achieved in many beacons, and the cost of stability in the receiver could be kept reasonably low.

This has another aspect: The bandwidth generally agreed upon in direction-finder receivers at present is many times larger than necessary, even for the purpose of identification of the radio beacon. The audio-frequency modulation of the beacon is not basically required for bearing indications. Provided sufficient stability is assured in the transmitter and in the direction-finder receiver, the present bandwidth of the receiver could be divided by 10; which is equivalent to an increase of 10 times the transmitter power. Much further improvement would still be possible, but might require costlier design. Further advantage can be derived from this increase of sensitivity by increasing the accuracy and stability of the direct-reading direction finder.

Very small bandwidths, of the order of 20 radio-frequency cycles, are being tried in connection with present developments of long-range navigational systems.

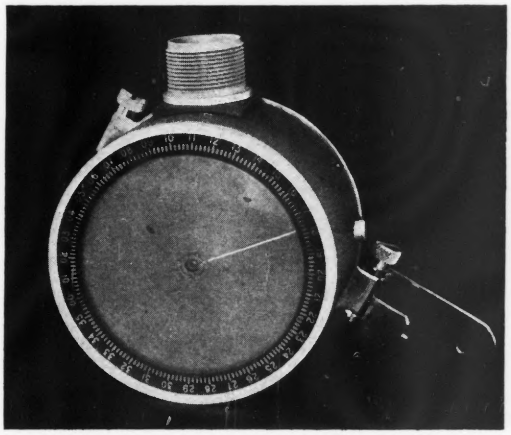
4.4 Pulse Transmission

Could night effect be reduced by the use of pulse transmission?

Pulse transmission for the purpose of discriminating between the ground wave and the sky wave was proposed and tested in Germany prior to the war. Expected results were obtained; the disadvantage being in the bandwidth required for the transmission. The circuits for separating the ground waves from the sky waves in the receiver would definitely be more expensive and

difficult to maintain than the simple circuits of the present receivers. Despite its promising aspects, this system would require a considerable amount of further development before a sound opinion could be formed concerning its practical value. It is still doubted that such an improvement, although resulting in increased range at night, would justify itself. However, the question would merit much more detailed consideration if advantage could be taken of the pulse transmission from the ground beacon for the purpose of distance measurement.

Any such modification of the present radio-beacon system on a world-wide basis would meet with difficulties of application that can readily be imagined, and these would contribute to discouraging the inventor or manufacturer.



A straight-line automatic bearing indicator applicable to most types of automatic direction finders.

4.5 Distance-Measuring System

Is the combination of the marine direction finder with some distance-measuring system desirable and/or possible?

In air navigation, the R-Theta system has reached justified success due to many advantages, two very important ones being:

A. A complete fix can be obtained from one station and supposedly from one transmission.

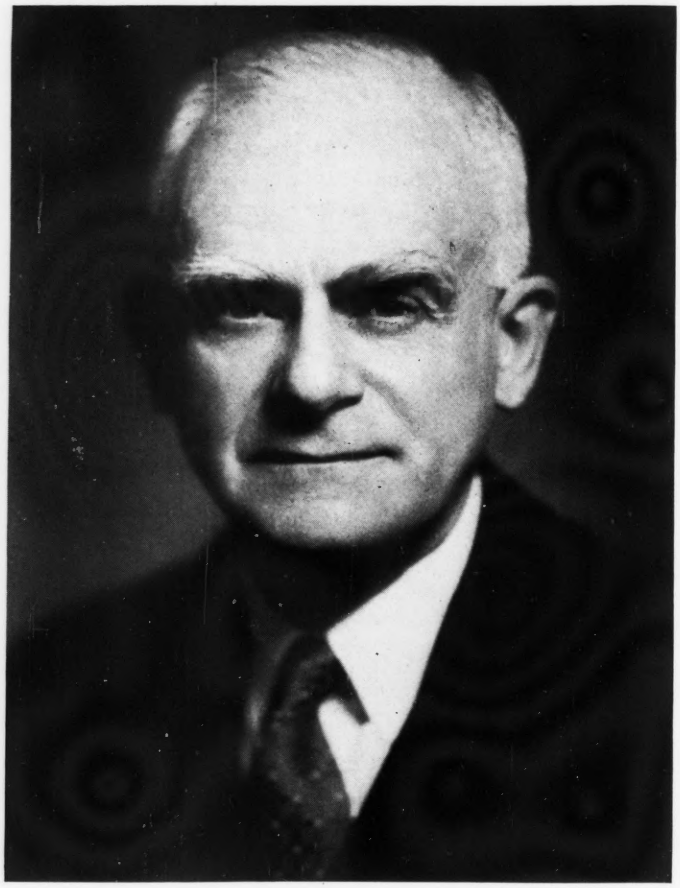
B. The accuracy of a fix is the same at all points in the 360-degree coverage of each station.

Distance-measuring equipment making use of ultra-high frequencies would not be of much assistance to the marine direction finder because its range is limited to line-of-sight. However, one can imagine a medium-frequency distance-measuring system using phase measurement of an audio-frequency modulation for the determination of time of travel, and, therefore, distance. The ship could interrogate the radio beacon on a medium-frequency channel assigned for that purpose; the audio-frequency signal received at the ground beacon would be retransmitted through the beacon to the ship, where the audio-frequency phase-displacement measurement, resulting in distance indication, would be effected at the output of the direction-finder receiver. The drawback to this system is that only one ship at a time could interrogate the beacon; but this interrogation should not need to last more than a number of seconds for each call, and could be limited in time. An indication of the availability of the channel for such interrogation could be provided, with self-identification of transmission. This distance-indicating system could be introduced in the form of attachments that would not require world-wide acceptance before advantages are realized from the first installations. Considerable developmental work would have to take place to solve such a problem properly, and it is very doubtful that a private enterprise would decide to start such a project unassisted.

5. Conclusions

Among the possible improvements to the marine direction finder that may be recognized in the not-too-distant future are automatic indication of bearing and a reconsideration of the stability and channeling of radio beacons. Moreover, it seems that the combination of the shipboard direction finder with some form of a distance-measuring system would result in an attractive navigational scheme, upon the development of which some imaginative effort could be applied in the next few years.

H. T. Kohlhaas, Editor, Retires



H. T. Kohlhaas has retired after twenty-three years as editor of *Electrical Communication*. Under his guidance, the magazine became a leading contemporary technical journal of communication development. It is among the leaders in circulation outside the U.S.A. Mr. Kohlhaas' retirement follows more than forty-two years with International Telephone and Telegraph Corporation and its predecessor companies. In addition to his editorship, he was assistant vice president of the Corporation and in charge of its Technical Publications Division.

From 1905 to 1922, Mr. Kohlhaas served in the Western Electric Company Laboratories (later Bell Telephone Laboratories) as an engineer, section head, and finally as a division head in the physical laboratory. In 1922, he was appointed executive personnel assistant to the systems engineer of the Bell Telephone Laboratories. He was transferred to the Western Electric Company headquarters in 1924.

In 1925, Mr. Kohlhaas was assigned to statistical studies and the editing of *Electrical Communication* for the International Western Electric Company. Through the purchase of that company, he became a member of the staff of the International Telephone and Telegraph Corporation. He edited *Electrical Communication* in London from 1937 through 1939 and, while there, served as publicity representative of the International Telephone and Telegraph Corporation and the International Standard Electric Corporation. He was appointed an assistant vice president of International Telephone and Telegraph Corporation in 1945.

Mr. Kohlhaas received a B.S. degree in electrical engineering from Cooper Union in 1907 and later was awarded electrical engineering degrees by that institution and Brooklyn Polytechnic Institute. Mr. Kohlhaas is a Member of the American Institute of Electrical Engineers and a Senior Member of the Institute of Radio Engineers. He is also a member of the Norwegian Chamber of Commerce.

Mr. Kohlhaas has agreed, at least for the present, to continue as consulting editor for *Electrical Communication*. This journal owes a great debt of gratitude to Mr. Kohlhaas and, while we wish him every enjoyment in his well-earned retirement, we also look forward to his continued interest and help as consultant.

FRANK C. PAGE

Contributors to This Issue

S. E. BUCKLEY

S. E. BUCKLEY was born in London on November 28, 1904. He received the B.Sc. degree from London University in 1924.

He joined the European engineering department of International Standard Electric Corporation in January, 1926, as a student engineer, and was transferred to the development and research laboratories at Hendon in 1929. He was again transferred in 1932 to Standard Telephones and Cables, Limited.

Associated with the development of Permalloy-type materials since 1927, he is at present responsible for development and engineering of magnetic materials including dust cores, loading coils, and telephone-cable accessories. Mr. Buckley is in charge of the chemical and metallurgical laboratories at the Woolwich plant.

HENRI BUSIGNIES

He is an Associate Member of the Institution of Electrical Engineers.

• • •

HENRI BUSIGNIES was born in Sceaux, France, on December 29, 1905. He received the degree of Electrical Engineer from Paris University in 1926.

On completion of his military service, he entered the Paris Laboratories of the International Telephone and Telegraph Corporation in 1928. Until about 1940, Mr. Busignies traveled to Italy, Spain, Switzerland, England, Africa, and the United States in the interests of the company. Since 1940, he has been an executive engineer and is now the Director of Federal Telecommunication Laboratories in New York city.

His first patent was issued in 1926 and has been followed by almost a hundred others. They have been dominantly in the field of automatic direction finding with particular attention to aircraft applications. He received the Lakhovsky award of the Radio Club of France in 1926.

Mr. Busignies is a Fellow of the Institute of Radio Engineers.

• • •

E. SERRA CALABUIG was born at Játiva (Valencia), Spain, in 1894. In 1910, he joined the State telegraph services. He continued his studies and obtained the degree of Ingeniero de Telecomunicación in 1920.

He became a plant engineer for the Compañía Telefónica Nacional de España in 1924 and is now director of construction and maintenance.

• • •

J. M. CLARA was born at Zaragoza, Spain, in 1892. He entered the State telegraph services in 1908 but continued his studies and received the degrees of Licenciado en Ciencias Químicas in 1914, Licenciado en Ciencias Físicas in 1916, Ingeniero de Telecomunicación in 1920, and in 1923 Ingeniero Radiotelegrafista ESE Paris.

In 1924, he joined the Compañía Telefónica Nacional de España as a district engineer and is now chief engineer. Since 1930, he has represented

E. SERRA CALABUIG

the Compañía Telefónica Nacional de España on the Comité Consultatif International Téléphonique.

• • •

THOMAS C. CLARK was born in Chicago, Illinois, on January 15, 1921. He received the B.S. degree in electrical engineering from Illinois Institute of Technology in 1942.

On leaving college, he accepted an engineering position with Federal Telephone and Radio Corporation and, from 1945 to 1947, worked on the development of airport traffic-control receivers. In October, 1947, he joined the engineering staff of Air Associates, Incorporated, in Teterboro, New Jersey.

J. M. CLARA

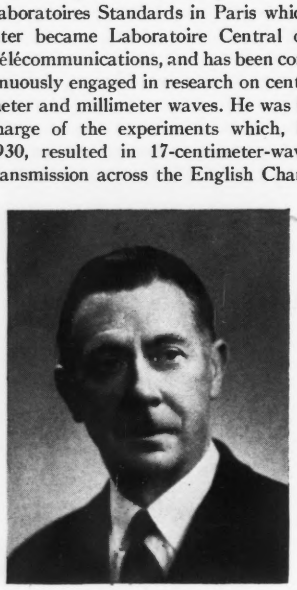
THOMAS C. CLARK

Mr. Clark holds an Associate membership in the Institute of Radio Engineers.

• • •

A. G. CLAVIER was born in Cambrai, France, in 1894. He received a degree in electrical engineering from Ecole Supérieure d'Electricité in 1919 and then joined the staff of engineers organized by General Ferrié at the Etablissement Central de la Radiotélégraphie Militaire. He was in charge of research on high frequencies from 1920 to 1925.

In 1929, Mr. Clavier joined Les Laboratoires Standards in Paris which later became Laboratoire Central de Télécommunications, and has been continuously engaged in research on centimeter and millimeter waves. He was in charge of the experiments which, in 1930, resulted in 17-centimeter-wave transmission across the English Channel and of the developments for the Lympne-St. Inglevert microwave radiotelephone link, which was inaugurated commercially in 1934. He was assistant director of research in 1945, when he was transferred to Federal Telecommunication Laboratories in New York, where he now holds the same position.



A. G. CLAVIER

He has published extensively on high-frequency oscillators, wave guides, and general electromagnetic theory, and has taught field theory and applications of ultra-high frequencies at the Ecole Supérieure d'Electricité.

Mr. Clavier is president of the section of the Société des Radioélectriciens dealing with hyperfrequencies. He is a Fellow of the Institute of Radio Engineers, and a Member of the Institution of Electrical Engineers.

• • •

EIGIL COHRT was born on November 19, 1902, in Copenhagen, Denmark. He studied at the Royal Technical College in Copenhagen from 1922 to 1926, when he joined the technical staff of the Bell Telephone Manufacturing Company in Antwerp.

Mr. Cohrt was transferred in 1938 to Standard Electric A/S in Copenhagen, where he became systems engineer and head of the circuit division.

• • •

ROBERT A. FELSENHELD was born on February 15, 1910, at East Orange, New Jersey. He has been active in the broadcast field since 1927. In 1941, he joined the staff of Federal Telecommunication Laboratories, Incorporated, to work on ultra-high-frequency antennas and allied subjects.

Mr. Felsenheld is a Member of the Institute of Radio Engineers.

• • •

SIDNEY FRANKEL was born on October 6, 1910 in New York City. Rensselaer Polytechnic Institute conferred three degrees on him: the B.A. degree in electrical engineering in 1931, and in mathematics the M.A. degree in 1934 and the Ph.D. degree in 1936. He was an instructor in mathematics from 1931 to 1933.

For a year after leaving college, Dr. Frankel served as a sound-recording engineer with the Brooklyn Vitaphone Corporation. In 1937–1938, he was an assistant engineer in the design and development of electronic flight instruments for the Eclipse Aviation Corporation.

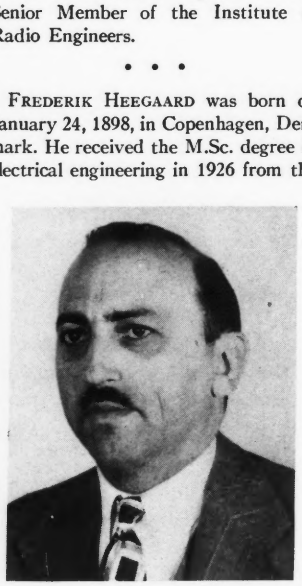
EIGIL COHRT

He joined the Federal Telegraph Company staff at Newark, New Jersey, in 1938 as an engineer on the design and development of radio transmitters. In 1943, he was transferred to Federal Telephone and Radio Laboratories, now Federal Telecommunication Laboratories. At present he is engaged in the development of components for microwave systems.

He is a member of Sigma Xi and a Senior Member of the Institute of Radio Engineers.

• • •

FREDERIK HEEGAARD was born on January 24, 1898, in Copenhagen, Denmark. He received the M.Sc. degree in electrical engineering in 1926 from the



ROBERT A. FELSENHELD

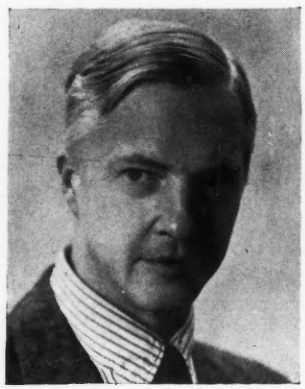
SIDNEY FRANKEL

Royal Technical College in Copenhagen. He then joined the engineering staff of the Post and Telegraph Administration and is now chief engineer of the broadcast service.

He served as vice chairman of the electrical section of the Society of Danish Engineers from 1935 to 1937, and as a member of its board of directors until 1941. He was chairman of the technical committee on Radio Broadcasting House. He is a member of the Academy of Technical Science and of its acoustical laboratory. He has been secretary of the Danish national committee for Union Internationale Radio Scientifique since 1938.

Mr. Heegaard was chief of the Danish delegation to the Bucharest radio conference in 1937, and a delegate to the International Telecommunication Conference in Cairo the following year and in Atlantic City in 1947. He was appointed a member of the Patentboard in 1942.

Knighthood in the Order of Danneborg was conferred on Mr. Heegaard in 1941.



FREDERIK HEEGAARD

• • •

HEINZ K. HENISCH was educated at Carlsbad, Czechoslovakia, and Ottershaw College, Chertsey, England. He then studied physics and mathematics at Bristol University and later at the University of Reading, where he obtained the degree of B.Sc. (First Class Gen. Hons.) in 1942.

He joined the radio department staff of the Royal Aircraft Establishment during the war, and is now again at Reading University, engaged on semiconductor research.

HEINZ K. HENISCH

Mr. Henisch became an Associate of the Institute of Physics in 1944.

• • •

ARMIG G. KANDOIAN was born on November 28, 1911 in Van, Armenia. In 1934, he received the B.S. degree and a year later the M.S. degree in electrical communication from Harvard University.

He joined the engineering staff of International Telephone and Telegraph Corporation in 1935. With its establishment in 1937, he was transferred to the International Telephone Development Company for work on air navigation and instrument landing systems. Federal Telephone and Radio Corporation took over the activities of that company and was succeeded by Federal Telecommunication Laboratories in 1945. Mr. Kandoian is now a department head responsible for radar, antennas, and multiplex broadcasting.

ARMIG G. KANDOIAN

Mr. Kandoian is a Senior Member of the Institute of Radio Engineers, an Associate of the American Institute of Electrical Engineers, and is a member of the Harvard Engineering Society and of Tau Beta Pi.

• • •

G. KLINGELFUSS was born in Aarau, Switzerland on October 19, 1895. After graduation from the technical school in Berthoud as an electrical engineer, he entered the Bell Telephone Manufacturing Company at Zurich in 1920.

In 1921, he was transferred to the engineering department in Antwerp,

G. KLINGELFUSS

ALBERT SHADOWITZ

and in 1923 assigned to sales engineering in the branch office at Berne. He served as commercial director in charge of the Swiss business from 1930. In 1941 the Berne branch was attached to Standard Telephone et Radio S.A., Zurich. Since 1935, he has been a member and secretary of the Board of Directors of the Zurich company.

• • •

ALBERT SHADOWITZ was born on May 5, 1915 at Brooklyn, N. Y. He received the B.E.E. degree from Polytechnic Institute of Brooklyn in 1935. After attending Harvard University during 1936 and 1937, he studied at the University of California from 1937 to 1940.

He served as an engineering employee at Aberdeen Proving Grounds from 1941 to 1943. In 1943, he joined Federal Telecommunication Laboratories as an engineer.

Mr. Shadowitz is an Associate Member of the Institute of Radio Engineers.

• • •

WILLIAM SICHAK was born on January 7, 1916, at Lyndora, Pennsylvania. He received the B.A. degree in Physics from Allegheny College in 1942.

From May, 1942, to November, 1945, he was engaged in developing microwave radar antennas at the Radiation Laboratory, Massachusetts Institute of Technology. Since November, 1945, he has been with Federal Telecommunication Laboratories, Incorporated, working on microwave antennas and allied subjects.

WILLIAM SICHAK

• • •

MARTIN SILVER graduated from New York University in 1941 with a Bachelor of Engineering degree. He then came to Federal Telecommunication Laboratories where he worked on communication-type jamming equipment. At the termination of the war, he was assigned to the development of frequency-modulation broadcast transmitters, monitors, links, and associated equipment.

Mr. Silver is an Associate Member of the Institute of Radio Engineers.

MARTIN SILVER